Exploiting phylogenetics to understand genome evolution in both modern and ancestral organisms

A Dissertation
Presented to
The Academic Faculty

By

Ziming Zhao

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology

August 2012

Exploiting phylogenetics to understand genome evolution in both modern and ancestral organisms

Approved by:

Dr. Eric A. Gaucher, Advisor
School of Biology
Georgia Institute of Technology

Dr. I. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. Hongwei Wu
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Frank J. Stewart
School of Biology
Georgia Institute of Technology

Date Approved: June 28, 2012

To my family and friends, for always supporting me…

ACKNOWLEDGEMENTS

I would like to thank my PhD advisor, Professor Eric A. Gaucher for all his long-term support and tremendous patience in mentoring my PhD research, revising my manuscripts, providing me with numerous opportunities to dive into the academic world, such as attending workshops and conferences, peer-reviewing scientific articles, conducting scientific outreach programs, and others. From my interactions with the Gaucher group, I am not only learning how to be a great enthusiastic scientist, but I am also learning how to shape my personality and build up my confidence to embrace my world ahead with a broader view, a bigger heart and a humble demeanor.

I would like to thank my committee members Dr. Soojin Yi, Dr. Frank Stewart, Dr. King Jordan, and Dr. Hongwei Wu, for contributing their time, efforts and scientific insights in supporting my PhD studies at Tech. I would like to thank my first mentor in graduate school Dr. Henry Wan, for his patience in guiding me into the scientific world and long-term support for my scientific career. I would like to thank my research collaborators Diego Fernando Mejia, Dr. Kevin Tang, Dr. Xianfu Wu, and Dr. Raul Perez Jimenez, for bringing together exciting projects to work on. I am thankful for tremendous assistance provided by graduate coordinators and administrative staff at Georgia Institute of Technology, with special thanks to Kevin Roman and Barbara Walker from School of Biology.

I would like to thank my lab mates from the Gaucher group, including Ryan Randall, James Kratzer, Joshua Stern, Dr. Megan Cole, Ercan Cacan, Benjamin Hsieh, and Dr. Betül Kacar, for all your support and all the beautiful memories we have shared in the

past years. Special thanks go to Dr. Kacar, for her inspiration, enthusiasm, and encouragement. I am thankful for all support and love given to me by my colleagues and friends at Georgia Institute of Technology, Center for Disease Control and Prevention and else where, including Jie Pan, Jianrong Wang, Thanawadee Preeprem, Linxin Gao, Hongjing Ma, Jing Cheng, Yuan Li, Mengwei Chen, Yao Wang, Chenyi Pan, Nan Hua, Xiaolei Mao, Lee Katzs, Ai-Ping Hu, Siwei Cao, Cong Feng, Siyuan Zhang, Ahsan Huda, Li Han, YunGyeong Lee, Yunho Jang, Hua (Angela) Yang, Chan Zhou, Hong Bo, Feng Liu, and others. To anyone I may have forgotten, please accept my genuinely apology and my sincere gratefulness.

Lastly, I would like to thank my beloved parents, sister, and other family members for their perpetual and absolute love and trust.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ..... **iv**

LIST OF TABLES ..... **xii**

LIST OF FIGURES ..... **xiii**

LIST OF ABBREVIATIONS ..... **xv**

SUMMARY ..... **xviii**

CHAPTER 1: INTRODUCTION **1**

Phylogenetics **1**

Exploiting phylogenetics to understand gene duplication and innovation **3**

Exploiting phylogenetics to understand genotype assignments and evolutionary pathways based on reassortments **5**

Exploiting phylogenetics to reconstruct ancestral states **7**

Description of this dissertation **10**

CHAPTER 2: GENOTYPIC DIVERSITY OF H5N1 HIGHLY PATHOGENIC AVIAN INFLUENZA VIRUSES IN ESTERN ASIA BETWEEN 1996 AND 2007 [31] **12**

Abstract **12**

Introduction **12**

Methods **15**

Viral RNA preparation and nucleic acid sequencing 15

Datasets 15

Progenitor gene identification 15
Phylogenetic analyses..... 16
Results 16
Diverse origins for gene segments of H5N1 HPAIVs 16
Combination of progenitor genes 27
Relationships of H5N1-PRs with reported reassortants 29
Emergence of low pathogenic AIVs through progenitor gene combinations 29
Discussion 29
Abbreviation 32
Acknowledgements 32
CHAPTER 3: EVOLUTION OF H5N1 HIGHLY PATHOGENIC AVIAN INFLUENZA VIRUSES IN VIETNAM BETWEEN 2001 AND 2007 [76] 34
Abstract 34
Introduction 34
Methods 34
Datasets and phylogenetic analyses. 35
Results and discussion 35
Emergence of H5 HA genes of AIVs in Vietnam 35
Phylogenetic Analyses of NA and Internal Genes Revealed an Abundant Genome Segment Pool in Vietnam 37
Abbreviation 40
Acknowledgements 40

CHAPTER 4: THE EVOLUTIONARY HISTORY OF THE CATENIN GENE FAMILY DURING METAZOAN EVOLUTION [17] 41

Abstract 41

Introduction 42

Methods 44

Gene presence/absence for the catenin family 44

Datasets 45

Phylogenetic analyses and tertiary structures 45

Annotation validation 46

Tissue specific gene expression 46

Results 46

Origins of the catenin subfamily members during metazoan evolution 46

Evolution of an ancient duplication in the catenin family 49

Origins and duplications of p120 subfamily members 51

Origins and duplications of beta subfamily members 54

Origins and duplications of alpha subfamily members 58

Annotation issues for the catenin family 59

Distant relatives of catenins in non-metazoans 63

Discussion 64

Origins and evolution for the catenin family 64

Evolution-related physiology of the catenin family 67

Conclusion 71

Abbreviation 71

Acknowledgments **72**

CHAPTER 5: ANCESTRAL SEQUENCE RECONSTRUCTION OF THIOREDOXIN TO UNDERSTAND ANCIENT ENVIRONMENTS [121] **73**

Abstract **73**

Introduction **73**

Methods **75**

Phylogenetic analyses 75

Ancestral sequence reconstruction 75

Results and discussion **75**

Phylogenetic tree of thioredoxin 75

Reconstruction of ancestral Trx enzymes 78

Conclusion **80**

Abbreviation **80**

Acknowledgments **81**

CHAPTER 6: ANCESTRAL GENOME RECONSTRUCTION AND MINIMAL GENOME OF MYCOPLASMAS **82**

Abstract **82**

Introduction **82**

Methods **89**

Data sources and phylogenetic analyses 89

Ancestral genome content reconstruction: 90

Comparing genome content of the mycoides cluster ancestor (MCA) with two hypothetical minimal genomes of mycoplasma 92

Functional annotations of gene lists: 93

Ancestral genome rearrangement reconstruction: 93

Results and discussion 93

Mycoplasma species tree and the search of outgroup 93

Ancestral genome content reconstruction 95

Comparing genome content of the mycoides cluster ancestor (MCA) with two hypothetical minimal genomes of mycoplasma 98

Genome rearrangement reconstruction 102

Conclusion 104

Abbreviation 105

Acknowledgements 105

CHAPTER 7: COMPUTATIONAL VALIDATIONS OF ANCESTRAL SEQUENCE RECONSTRUCTION 107

Abstract 107

Introduction 107

Methods 109

Phylogeny of LeuB in *Bacillus* 109

Computational Simulation of ancestral sequence reconstruction 110

Ancestral sequence reconstruction of LeuB 110

Results and discussion 111

Validation of LeuB phylogenetic tree used in Hobbs' paper 111

Computational Simulation of ancestral sequence reconstruction 114

Ancestral sequence reconstruction of LeuB 115

Conclusion **117**

Abbreviation **117**

Acknowledgements **117**

CHAPTER 8: CONCLUSION **119**

APPENDIX A: SUPPLEMENTARY INFORMATION FOR CHAPTER 2 ..... **124**

APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER 3 ..... **142**

APPENDIX C: SUPPLEMENTARY INFORMATION FOR CHAPTER 4 ..... **150**

APPENDIX D: SUPPLEMENTARY INFORMATION FOR CHAPTER 5 ..... **156**

APPENDIX E: SUPPLEMENTARY INFORMATION FOR CHAPTER 6 ..... **160**

APPENDIX F: SUPPLEMENTARY INFORMATION FOR CHAPTER 7 ..... **177**

PUBLICATIONS **183**

REFERENCES **185**

LIST OF TABLES

TABLE 2.1: POTENTIAL PROGENITOR GENES IDENTIFIED FOR H5N1 HIGHLY PATHOGENIC AIVS. 26

TABLE 2.2: THE REASSORTANTS FROM COMBINATIONS OF PROGENITOR GENES FOR H5N1 HPAIVS AND SOME LOW PATHOGENIC AIVS IDENTIFIED IN EASTERN ASIA.....28

SUPPLEMENTARY TABLE A.1: REASSORTANTS GENERATED FROM THE PUTATIVE PRECURSOR GENES IDENTIFIED IN THIS STUDY. 132

SUPPLEMENTARY TABLE C.1: SUMMARY OF ANNOTATION PROBLEMS FOR THE CATENIN FAMILY. 153

SUPPLEMENTARY NOTE D.1: THIOREDOXIN SEQUENCES USED FOR ANCESTRAL SEQUENCES RECONSTRUCTION. 157

SUPPLEMENTARY TABLE E.1: LIST OF 668 GENES IN THE ANCESTOR OF THE MYCOIDES CLUSTER (MCA – MYCOIDES CLUSTER ANCESTOR).....160

SUPPLEMENTARY TABLE E.2: LIST OF 41 GENES PRESENT IN THE MYCOIDES CLUSTER ANCESTOR (MCA) AND THE PROTEOME OF *M. GENITALIUM* (MG476), BUT ABSENT IN THE MINIMAL GENOME OF *M. GENITALIUM* (MG381).....172

SUPPLEMENTARY TABLE E.3: LIST OF 105 GENES PRESENT IN MCA AND THE PROTEOME OF M. PULMONIS, BUT ABSENT IN THE MINIMAL GENOME OF M. PULMONIS (MP310).....173

SUPPLEMENTARY TABLE E.4: LIST OF 23 GENES IN THE MCA ANCESTOR, BUT ABSENT IN THE HYPOTHETICAL MINIMAL GENOMES OF EITHER *M. PULMONIS* OR *M. GENITALIUM*.....176

LIST OF FIGURES

FIGURE 2.1: PHYLOGENETIC TREES OF HA(A), NA (B), PB2(C), PB1(D), PA(E), NP(F), MP(G), AND NS (H) FOR H5 HPAIVS AND ASSOCIATED PROGENITOR GENES. 18

FIGURE 3.1: PHYLOGENETIC TREE OF SUBTYPE H5 HA GENES FROM AVIAN INFLUENZA VIRUSES. 37

FIGURE 3.2: PHYLOGENETIC TREE OF THE NA GENE OF HPAIV H5N1 VIRUSES. 39

FIGURE 4.1: THE PRESENCE/ABSENCE OF CATENIN FAMILY MEMBERS IN SPECIES DURING METAZOAN EVOLUTION. 48

FIGURE 4.2: THE EVOLUTION OF THE CATENIN FAMILY. 51

FIGURE 4.3: THE EVOLUTION OF THE P120 SUBFAMILY. 53

FIGURE 4.4: THE EVOLUTION OF THE BETA CATENIN SUBFAMILY. 57

FIGURE 5.1: PHYLOGENETIC TREE USED FOR THE ANCESTRAL SEQUENCE RECONSTRUCTION OF TRX ENZYMES. 77

FIGURE 5.2: PHYLOGENETIC ANALYSIS OF TRX ENZYMES AND ANCESTRAL SEQUENCE RECONSTRUCTION. 79

FIGURE 6.1: THE EVOLUTION AND TAXONOMY CLASSIFICATION OF MOLLICUTES. 85

FIGURE 6.2: GENOME EVOLUTION AND ANCESTRAL GENOME RECONSTRUCTION IN SMALL- AND LARGE-SCALES. 88

FIGURE 6.3: THE FLOWCHART FOR GENOME CONTENT RECONSTRUCTION. 92

FIGURE 6.4: GENOME CONTENT RECONSTRUCTION AND FUNCTIONAL DISTRIBUTION OF GENES UNIQUE IN THE MYCOIDES CLUSTER. 98

FIGURE 6.5: FLOWCHARTS OF THE PROCEDURES AND RESULTS FOR GENOME CONTENT COMPARISON OF THE MYCOIDES CLUSTER ANCESTOR (MCA) AND TWO HYPOTHETICAL MINIMAL GENOMES OF *M. GENITALIUM* (MG) AND *M. PULMONIS* (MP). 101

FIGURE 6.6: MULTIPLE GENOME ALIGNMENT AND GENOME REARRANGEMENT RECONSTRUCTION OF THE MYCOIDES CLUSTER. 103

FIGURE 7.1: THE VALIDATION OF THE TREE TOPOLOGY OF LEUB UTILIZED BY HOBBS' WITH BRANCHES SUPPORTED WITH BOOTSTRAP VALUES FROM GARLI AND POSTERIOR PROBABILITY VALUES FROM MRBAYES.....113

FIGURE 7.2: ACCURACY AND CORRECTNESS OF ASR BY COMPUTATIONAL SIMULATIONS. 115

SUPPLEMENTARY FIGURE A.1: PHYLOGENETIC TREES OF PB2, PB1, PA, HA, NP, NA, MP, AND NS FOR H5 HPAIVS AND ASSOCIATED PROGENITOR GENES.....124

SUPPLEMENTARY FIGURE B.1: PHYLOGENETIC ANALYSES OF THE H5N1 HIGHLY PATHOGENIC AVIAN INFLUENZA VIRUSES (HPAIVS) ISOLATED IN VIETNAM BETWEEN 2001 AND 2007.....142

SUPPLEMENTARY FIGURE B.2: HK821-LIKE VIRUSES FORMED THREE SUB-LINEAGES: HK821P, HK821α, AND HK821β.....149

SUPPLEMENTARY FIGURE C.1: BAYESIAN PHYLOGENIES FOR THREE SUBFAMILIES OF THE CATENIN FAMILY. 150

SUPPLEMENTARY FIGURE F.1: MAXIMUM LIKELIHOOD AND BAYESIAN PHYLOGENIES OF LEUB PROTEINS BY GARLI AND MRBAYES RESPECTIVELY.....177

SUPPLEMENTARY FIGURE F.2: ACCURACY AND CORRECTNESS OF ANCESTRAL SEQUENCE RECONSTRUCTION BASED ON COMPUTATIONAL SIMULATIONS.....179

SUPPLEMENTARY FIGURE F.3: MULTIPLE SEQUENCE ALIGNMENT OF COMPUTATIONALLY RECONSTRUCTED LEUB ANCESTRAL SEQUENCES BY US USING DIFFERENT MODELS AND DATASETS AND BY HOBBS. 180

LIST OF ABBREVIATIONS

GTR	Generalized time reversible
dN	The number of non-synonymous substitutions
dS	The number of synonymous substitutions
aa	Amino acid
JC69	Jukes and Cantor, 1969
K80	Kimura, 1980
F81	Felsenstein, 1981
HKY85	Hasegawa-Kishino-Yano, 1985
BLAST	Basic local alignment search tool
LRT	Likelihood ratio test
AIC	Akaike information criterion
BIC	Bayesian information criterion
MCMC	Markov chain Monte Carlo
UPGMA	Unweighted Paired Group Method with Arithmetic Means
ME	Minimal Evolution
ML	Maximum Likelihood
NJ	Neighborhood Joining
MP	Maximum Parsimony
HPAIVs	Highly Pathogenic Avian Influenza Viruses
AIVs	Avian Influenza Viruses
HA	Hemagglutinin
NA	Neuraminidase

NP	Nucleoprotein
PR	Progenitor Reassortant
GIN	Genotype In Network
NCVD	National Center for Veterinary Diagnostics
CCV	Complete Composition Vector
ESS	Effective Sample Size
GIN	Genotype In Network.
NCVD	National Center for Veterinary Diagnostics
P120	p120 catenin
ARVCF	Armadillo Repeat protein deleted in Velo-Cardio-Facial syndrome
pkp	Plakophilin
PP	Posterior Probability
ARM	Armadillo
PSRF	Potential Scale Reduction Factor
ASR	Ancestral Sequence Reconstruction
Myr	Millions years
Gyr	Billion years
EF-Tu	Elongation factor thermo unstable
LACA	Last archaeal common ancestor
AECA	archaea/eukaryota common ancestor
Trx	thioredoxin
M-PAS	Most probabilistic ancestral sequence
AGR	Ancestral Genome Reconstruction
Trp	Tryptophan
TBR	Tree bisection reconnection

MCA	Mycoides Cluster Ancestor
MSC	*Mycoplasma mycoides mycoides*
MLC	*Mycoplasma mycoides capricolum*
MCAP	*Mycoplasma capricolum*
MSB	*Mycoplasma leachii*
Mf	*Mesoplasma florum*
Mp	*Mycoplasma pulmonis*
Mg	*Mycoplasma genitalium*
AYWB	*Phytoplasma AYWB*
PAM	*Phytoplasma OY-M*
bp	Base pair
COG	Cluster of Orthologous Groups
KAAS	KEGG Automatic Annotation Server
PTS	Phosphotransferase system
BBH	Bidirectional Best Hit
LeuB	3-isopropylmalate dehydrogenase
PACE	The Partnership for an Advanced Computing Environment

SUMMARY

In this dissertation, computational evolutionary analyses, particularly phylogenetics and ancestral reconstruction, have been extensively exploited to better understand both functional divergence within individual gene families on the small-scale as well as gene content/organization at the genomic level on the large-scale. These small-scale studies focus on two gene families, thioredoxin and catenin, intended to deepen our understanding of both protein adaptation and innovation of new gene families through duplication events, respectively. Alternatively, the large-scale studies focus on both reassortments as revealed by diverse genotypes of H5N1 avian influenza viruses as well as inferences of gene content and genome rearrangements as revealed by ancestral genome reconstruction of a hypothetical ancient Mycoplasma species.

Such evolutionary studies provide us with insights into biological phenomena that in turn can be exploited for different purposes. For instance, studies of viral epidemics and modes of transmission by assigning genotypes of H5N1 highly pathogenic avian influenza viruses can help us to better prepare, prevent and control diseases. Determining functional divergence following an array of duplications within a cancer-related catenin gene family improves our understanding of developmental physiology within Metazoans. Resurrected ancient thioredoxin proteins based on computational ancestral sequence reconstruction provide possible clues to the environments that hosted early life. A set of genes inferred computationally to compose an ancient genome using Mycoplasma allow us to link genotypes with lineage-specific phenotypes and also facilitate synthetic

biology's attempt to create a viable, self-sustainable organism consisting of a recombinant, minimal genome.

Beyond case studies of natural evolution, this dissertation also describes my efforts to better understand methods of ancestral sequence reconstruction. Such work consisted of computational analysis of an experimentally-derived data set in order to benchmark these methods as well as conducting simulations. In total, this particular computational work provides us with greater insights to the accuracies and limitations of ancestral sequence reconstruction methods.

The work presented in this dissertation highlights the diverse questions that evolutionary studies attempt to address and the different biological levels that can be studied to answer these questions.

CHAPTER 1

INTRODUCTION

Nothing in biology makes sense except in the light of evolution. - Th. Dobzhansky, 1973

Nothing in evolutionary biology makes sense except in the light of a phylogeny. - J. M. Savage, 1997

Phylogenetics

Phylogenetics is used to study evolutionary relationships among biological entities (genes or species), often by using molecular data to recapitulate evolutionary events such as speciation, gene duplication and horizontal gene transfer, thus providing insight into the rise of orthologs, paralogs and xenologs [1]. Various evolutionary models can be incorporated for phylogenetic analysis using molecular data, and model selection can be used to determine an optimal statistical evolutionary model among many. An optimal model can be selected based on different statistical criteria such as Likelihood Ratio Test (LRT), Akaike information criterion (AIC), and Bayesian information criterion (BIC) [2]. Software packages have been developed for model selection, such as jModeltest [2] and MrModeltest [3] for DNA sequences, and ProtTest for protein sequences [4, 5].

Phylogenies can be inferred using different algorithms that include distance-based approaches such as Unweighted Paired Group Method with Arithmetic Means (UPGMA), minimum evolution (ME), least square, and neighbor-joining (NJ) algorithms or include character-based approaches such as maximum likelihood (ML), maximum parsimony (MP), and Bayesian algorithms [6]. Different algorithms have been discussed

for their strengths and weaknesses in a recent review by Ziheng Yang [6]. Briefly, distance-based approaches create a distance matrix first by pairwise sequence comparisons, and then a phylogenetic tree is constructed according to the pairwise distance matrix by using an algorithm that minimizes the branch distances among all the pairwise comparisons [6]. Alternatively, character-based approaches consider each character (e.g., DNA or aa sites) in a multiple sequence alignment in order to calculate a 'tree score', and the 'tree score' refers to the minimum number of changes in MP, a likelihood score in ML, and a posterior probability in Bayesian approaches [6]. Generally speaking, distance-based approaches are very fast due to the clustering algorithm and can thus be applied to large dataset. However, distance-based approaches have a difficult time fitting sequences that are highly diverged or sequences with many gaps. MP is simple and easy to implement, however, it does not incorporate evolutionary models and thus suffers from long-branch-attraction and other well-known confounding issues. Both ML and Bayesian methods can incorporate sophisticated evolutionary models and are capable of handling more divergent sequences. In addition, ML generates likelihood scores and thus can be scrutinized using a likelihood ratio test to select a best-fitting model for the chosen sequences. Phylogenetic trees can be evaluated by the bootstrapping procedure - with a 70% bootstrap value used as a cutoff corresponding to 95% statistical confidence under certain assumptions related to tree symmetry, evolutionary rate and sequence divergence [7]. By contrast, Bayesian methods are attractive by not only incorporating prior distributions of parameters, but also utilizing Markov Chain Monte Carlo (MCMC) algorithms for computational speed, and a posterior probability of a tree can be directly

explained as the confidence of the tree, albeit, posterior probabilities can be overestimated [6].

Briefly, phylogenetics has been widely used by biologists for diverse applications such as annotating genes and taxonomic classifications [8], identifying criminals by monitoring HIV transmission based on molecular data in forensic science [9], and triggering the emergence of the new discipline phylomedicine by its integration with genomic medicine [10].

Exploiting phylogenetics to understand gene duplication and innovation

Gene duplication, first documented by its discovery in the fruit fly in 1930s and widely recognized by its role in evolution as summarized by Susumu Ohno's seminal book '*Evolution by Gene Duplication*' in 1970s, has been widely observed in all three domains of life [11]. Gene duplication, from the phrase itself, is a duplication of a gene in a DNA region, and the molecular mechanisms responsible for gene duplication include homologous recombination, retrotransposition, or whole chromosome/genome duplication [11].

Phylogenetics can be used to infer evolutionary origins and relationships of duplicated genes (paralogs), and distinguish gene duplication from other evolutionary mechanisms such as horizontal gene transfer that gives rise to xenologs or speciation that gives rise to orthologs. At the heart of such analyses, of course, a species tree is required for such resolution along with a gene tree [12]. Individual or concatenated orthologs are generally used for inferring a species tree; these orthologs can be genes that are highly conserved such as 16S rRNA, or genes involved in replication, transcription and translation, concatenated orthologs can be achieved by concatenating core genes that are

present in all interested species based on pairwise genome content comparisons [13]. Species tree can be also achieved based on large-scale genome evolution using a pairwise distance matrix based on gene presence and absence status derived from genomic content [14], and/or genomic rearrangements [15, 16]. A gene tree is constructed based on sequences retrieved from an exhaustive species-specific search for paralogs and an extensive search for orthologs in representative species across the species tree; the exhaustive search of sequences within a particular species and across a species tree can be based on sequence similarity using tools like BLAST [17]. Based on both the gene tree and the species tree, gene duplications occurring at different levels (species-specific, phylum-specific, etc.) can be inferred by having a bifurcation of gene members from individual monophyletic clades [17]. Accordingly, gene duplication has been utilized to infer the root of the tree of life at a position between last common ancestor of bacteria and the last common ancestor of eukaryotes and archaea by tracing the connection of the bifurcation that connects two subtrees using an ancient duplication that occurred before the divergence of three domains of life [18]. Additionally, evolutionary relationships inferred from duplicated genes can help resolve gene annotation for gene members within a gene family, as a complement to sequence similarity search [17].

Evolutionary fates of duplicated genes consist of nonfunctionalization (pseudogenization), subfunctionalization (preservation of partial ancestral functions), and neofunctionalization (derivation of novel functions); hence gene duplication plays a major role in genome expansion and innovation [11, 19]. Functional divergence resulting from subfunctionalization and neofunctionalization after a species or lineage/cluster (e.g. phylum) can be understood on the molecular level due to relaxed or shifted selective

constraints following gene duplication. Functional divergence sites, categorized into type I (shifted evolutionary rates) and type II (shifted amino acid property or constant-but-different), can be inferred based on cluster-specific protein sequence comparisons along a phylogenetic tree [20].

Exploiting phylogenetics to understand genotype assignments and evolutionary pathways based on reassortments

Reassortment, antigenic shift and antigenic drift are terms describing genomic evolution on both large- and small-scales, particularly for influenza viruses. Here, reassortment refers to a large-scale evolutionary event caused by recombination of the eight individual gene segments composing a viral genome while antigenic shift refers to reassortment of gene segments restricted to two antigenic surface proteins hemagglutinin (HA) and neuraminidase (NA) as opposed to the eight gene segments in the genome. Antigenic drift refers to small-scale genome evolution with sequence substitutions occurring at antigenic sites of HA and NA that enable viruses to evade host immune systems such as antibody binding [21].

Influenza viruses are single stranded, negative sense, RNA viruses from the family Orthomyxoviridae infecting birds and mammals [22, 23]. Influenza viruses are classified into type A, B, and C, and influenza A virus, naturally hosted by birds, is the most virulent type and caused three flu pandemics in the 20^{th} century [23]. Influenza A viruses can be further divided based on serotypes of two antigens HA and NA, with 16 HA subtypes and 9 NA subtypes [23-25]. Particularly, my work focuses on understanding the emergence and transmission pathways of highly pathogenic H5N1 avian influenza viruses circulating in East Asia from its first isolation in 1996 to 2007.

Phylogenetic analysis of these highly pathogenic H5N1 virus strains from 1996 to 2007, together with influenza A viruses circulating before 1996, was performed for all eight gene segments independently based on genome evolution from both reassortment and small-scale sequence mutations. When two viruses infect the same host cell, the genomes that comprise the eight separate gene segments can reassort and assemble a new virus strain due to the emergence of a novel genotype. Phylogenetics can be used to trace such reassortment events for each gene segment by assigning a precursor or ancestor for each pathogenic H5N1 virus between 1996 and 2007 using non-pathogenic influenza A viruses circulating before 1996, thus each virus can be represented by an individual genotype based on a combination of eight precursors for eight gene segments independently [25]. The total number of distinct genotypes of H5N1 viruses circulating between 1996 and 2007 can be then summarized by gathering each genotype achieved from individual viral strains, and the evolutionary origins of each virus can be tracked easily by its genotype composed of eight precursor genes. Accordingly, each unique genotype represents a series of viruses isolated from a range of years and various geographic locations, thus modes of viral transmission can be retraced based on genotypes, viral isolation time and geographic locations. Therefore, genotype assignments based on reassortment can help us determine the origins and patterns of transmission of highly pathogenic H5N1 avian influenza viruses.

The importance of genotype assignments based on reassortment for influenza viruses is due to its association with virus pathogenicity (genotypes of A, B and E), host specificity (wider host range associated with specific genotypes), and transmissibility (more efficient for specific genotypes) [26]. A variety of small-scale mutations have been

also reported for their association with virus pathogenicity (a polybasic amino acid insertion at the cleavage site of HA, glycosylation patterns of HA, and a 22 amino acid deletion in the NA stalk) [27] and host specificity (lysine at position 627 instead of glutamic acid for polymerase PB2 is associated with host shift from avian to humans) [28].

In summary, phylogenomic analyses that assign genotypes and that map small-scale mutations of H5N1 viruses from 1996 to 2007 can provide insights into understanding viral pathogenicity, host specificity, and transmissibility to facilitate virus surveillance and vaccine designs, and better prepare, prevent and control flu pandemics [29]. The emergence of novel genotypes and patterns of transmission of these influenza viruses, together with information of virus isolation time and geographic locations provides insights into the mechanisms of the birth of novel genotypes under particular ecological conditions and predictions of modes of viral transmission by considering paths of migratory birds [30, 31].

Exploiting phylogenetics to reconstruct ancestral states

Ancestral gene resurrection, first proposed by Pauling and Zuckerkandl in 1963 [32] and slowly developed along with DNA synthesis techniques in the 1990s, enables us to travel back in time and infer ancient molecules [33]. Ancestral gene resurrection incorporates computational reconstruction of ancestral sequences and experimental synthesis and expression of ancient genes in the laboratory [33]. Studies of biochemical functions of resurrected genes such as steroid hormone receptors, alcohol dehydrogenases, elongation factor thermo unstable (EF-Tu), and thioredoxin provide us with valuable information regarding paleogenetics and paleobiochemistry by

uncovering ‘genetic footprints’ from genetic information maintained in present-day organisms [34-39].

Ancestral sequence reconstruction (ASR) can be achieved computationally using a multiple sequence alignment, phylogeny construction, and ancestral state inference. Different algorithms (ML, MP and Bayesian methods) are implemented in ASR, and the strength and weakness of different algorithms in ASR are similar to those in phylogenetic tree inference discussed in the previous section [33]. MP was initially implemented in ASR by Steven Benner’s group in 1990s for the ancestral resurrection of digestive ribonuclease from artiodactyls [40], and the accuracy of MP in ASR has been experimentally tested by David Hillis and his colleagues in their experimental phylogeny work using viruses [41]. Similar to its limitations in phylogeny reconstruction, MP is not sufficient with divergent sequences and it lacks an evolutionary model [33]. In contrast, ML incorporates explicit evolutionary models and can handle more divergent sequences [33], and ML was first implemented for ASR in the PAML software package by Ziheng Yang in 1995 [42]. The strength of ML is also due to the incorporation of parameters and models that fit the data and the incorporation of a Bayesian posterior probability for the ancestral state inference at each site in order to evaluate the accuracy of the inferred states [33]. A computational simulation published by Richard Goldstein’s group in 2006, however, indicates both MP and ML methods are biased towards overestimating properties such as thermostability when inferring ancestral states, while the Bayesian approach performs better despite choosing a less-probable ancestral state from the posterior probability distribution [43]. However, another computational simulation published by Joseph Thornton’s group in 2010 indicates that the Bayesian method that

incorporates phylogeny uncertainties is not better than the ML method in the accuracy and robustness of ancestral inferences since the uncertainties in phylogeny construction apply to uncertainties in ancestral inference as well [44]. In summary, computational ASR is a powerful tool to uncover evolutionary pathways and infer properties of ancient molecules [38, 39]. Ancestral resurrection, of course, largely relies on accurate computational inferences of ancestral sequences, thus incorrect sequences can lead to incorrect biological conclusions [33]. Therefore, computational ASR must be performed carefully with correctly aligned sequences, a (nearly) correct tree topology, and an optimal evolutionary model in order to have confidence with the experimental synthesis, expression and biochemical studies of ancient proteins that attempt to answer interesting biological questions.

In this regard, ASR could benefit from additional studies that resurrect different gene families to determine whether results from additional studies are consistent with previously published studies. This would not directly add rigor to the ASR field, but consistency among different studies is an important aspect of this burgeoning field. In another regard, ASR could benefit from additional computational analyses that attempt to further understand the specific performances of different ASR algorithms under different phylogenetic conditions. Such computational studies would directly add rigor to the field.

I have attempted to address both of these regards with my research. Along these lines, I have resurrected a gene family that will allow us to determine if inferences regarding paleoenvironments based on ancient proteins from this gene family are consistent with previous ASR studies from the Gaucher group. Specifically, I intend to determine whether the paleoenvironment from the Precambrian era that hosted ancient life was hot

based on the biochemical behaviors of ancient thioredoxin genes resurrected from a large phylogenetic distribution. The trends for temperature and pH preference were determined by resurrecting thioredoxin at a series of ancestral nodes that included the last bacterial common ancestor, the last eukaryotes common ancestor, the last archaea common ancestor, and others.

Along another line, I have attempted to validate the accuracy and correctness of computational reconstruction methods by testing different algorithms such as maximum parsimony and maximum likelihood using various evolutionary models, different types of datasets (DNA, codons and amino acids), and different phylogenetic topologies. I have utilized both computer-simulated datasets and experimentally-derived datasets (although the latter is not discussed here). I have determined how the accuracy and correctness of computational sequence reconstruction can be biased due to several factors. In total, this particular computational work provides us with greater insights to the accuracies and limitations of computational sequence reconstruction methods in general.

Description of this dissertation

My dissertation focuses on exploiting phylogenetics using different algorithms and evolutionary models to better understand genome evolution from both small- and large-scale perspectives in order to assign genotypes based on assortment, resolve species relationships and gene annotation issues, further understand gene gain/loss within individual gene families, measure functional divergence among homologs, and infer ancestral character states.

Chapter Two and Three describe my research on genomic evolutionary studies of highly pathogenic H5N1 avian influenza viruses circulating in East Asia from 1996 to

2007 in order to determine evolutionary origins and modes of transmission of these viruses for purposes of disease epidemic/pandemic preparation, prevention and control. Chapter Four describes my research on the evolutionary history of the gene family catenin in order to determine evolutionary patterns and functional divergence of this gene's role in multicellularity and developmental physiology. Chapter Five describes my research on reconstructing the ancestral sequences of a ubiquitous enzyme called thioredoxin in order confirm (or refute) previous studies that examined the potential environment of life on early Earth. Chapter Six describes my research on reconstructing the ancestral genome of Mycoplasma in order to understand genome evolution and facilitate minimal genome studies in the era of synthetic biology. Chapter Seven describes my research that attempts to validate ancestral sequence reconstruction (ASR) using computational simulations.

The diverse evolutionary studies presented in this dissertation deepen our fundamental understanding of multiple perspectives of biology and highlight the importance of using evolutionary analyses to answer diverse biological questions.

CHAPTER 2

GENOTYPIC DIVERSITY OF H5N1 HIGHLY PATHOGENIC AVIAN INFLUENZA VIRUSES IN ESTERN ASIA BETWEEN 1996 AND 2007 [31]

Abstract

Besides enormous economic losses to the poultry industry, H5N1 highly pathogenic avian influenza viruses (HPAIVs) originating in eastern Asia have posed serious threats to public health. Up to April 17, 2008, 381 human cases had been confirmed with a mortality of more than 60 %. Here, we attempt to identify potential progenitor genes for H5N1 HPAIVs since their first recognition in 1996; most were detected in the Eurasian landmass before 1996. Combinations among these progenitor genes generated at least 21 reassortants (named H5N1 progenitor reassortant, H5N1-PR1–21). H5N1-PR1 includes A/Goose/Guangdong/1/1996(H5N1). Only reassortants H5N1-PR2 and H5N1-PR7 were associated with confirmed human cases, H5N1-PR2 in the Hong Kong H5N1 outbreak in 1997 and H5N1-PR7 in laboratory confirmed human cases since 2003. H5N1-PR7 also contains a majority of the H5N1 viruses causing avian influenza outbreaks in birds, including the first wave of genotype Z, Qinghai-like and Fujian-like virus lineages. Among the 21 reassortants identified, 13 are first reported by us. This study illustrates evolutionary patterns of H5N1 HPAIVs, which may be useful toward pandemic preparedness as well as avian influenza prevention and control.

Introduction

In 1996, two strains of HPAIVs, A/Goose/Guangdong/1/96 (H5N1) and A/Goose/Guangdong/2/96 (H5N1) (called GsGd viruses), were isolated from sick geese in Shanshui, a small middle-western town in Guangdong Province, China [45, 46]. About one year later, a related H5N1 genotype caused human deaths in Hong Kong, having been isolated first from chicken farms and later in the live poultry markets [47]. This Hong Kong 1997 (HK97) H5N1 highly pathogenic avian influenza outbreak was the first documented incident of a purely AIV causing human respiratory disease and death. After slaughtering 1.5 million chickens the avian influenza outbreak stopped and was followed by banning live poultry trade for 7 weeks [48, 49]. There have been no known indigenous human H5N1 cases in Hong Kong since the 1997 incident. In early 2003, a new genetic variant was isolated from two Hong Kong residents one of whom subsequently died soon after returning from a visit to Fujian Province, China to the north of Hong Kong (www.who.int/influenza/human_animal_interface/en/). This new H5N1 genetic variant led to avian influenza outbreaks in southeastern Asia at the end of the year. In May 2005, another H5N1 genetic variant was identified in Qinghai Lake, western China and spread to central Asia and beyond [50, 51]. Up to April 17, 2008, 381 human cases have been confirmed and 235 were fatal (www.who.int). At least 209 million birds have been slaughtered or died of this disease since 2003 (www.fao.org).

Southern China has an abundance and diversity of AIVs in domestic ducks which are raised in close proximity to humans [52] in a region designated as a hypothetical epicentre for the emergence of pandemic influenza viruses [53]. However, with such an abundance of AIVs, why don't influenza pandemics occur more frequently? There must be something very complex that is needed to cause one or more AIVs to give rise to a

pandemic virus. Therefore, by reconstructing genetic events that may have led to the appearance of HK97 and current H5N1 HPAIVs in humans, the information generated might provide a higher order of preparedness for a pandemic in the future.

The isolation of H5N1 viruses from chickens and humans in Hong Kong in 1997 possibly pre-empted a pandemic. The availability of viral sequences provides us with an opportunity to track down the genetic origins of these viruses. It has been reported that the HA of A/Goose/Guangdong/1/96 (H5N1) most likely provided the HA of HK97 H5N1 HPAIVs [54]. During the past decade, a range of H5N1 reassortants has been reported [26, 55-58]. Recently, segments from AIVs isolated 30 years ago were identified in these H5N1 AIVs [59, 60]. However, the genetic genesis of these H5N1 AIVs have not been well characterized yet due to i) lack of systematic surveillance in avian hosts in this region from 1980 to 2000, and ii) limitation of phylogenetic analysis approaches. A new quantitative genotype method called Genotype In Network (GIN) was developed [59]. Different from conventional phylogenetic tree construction approach, GIN does not perform multiple sequence alignment or tree construction and is able to analyze a large number of viruses. By combining with phylogenetic tree construction, this method provides an opportunity towards a more systematic genetic analysis of a wider range of viruses pre-and post-1996 [59].

In this study, we analyzed the genes of H5N1 HPAIVs from 1996 onward and attempted to identify possible progenitors for them. The results indicate that at least 21 reassortants have emerged from combinations of these genes since then. The information generated may be useful for pandemic preparedness as well as avian influenza prevention and control.

Methods

Viral RNA preparation and nucleic acid sequencing

The RNA for two archive influenza viruses, Tk/England/50-92/91 (H5N1) and Tk/England/N-28/73 (H5N2), was kindly provided by Dr. David Swayne USDA Southeastern Poultry Research Laboratory. Viral RNA was amplified using One-Step RT-PCR kit (QIAGEN). The amplified products were cleaned using USB EXOSAP-IT PCR products clean up kit (USBWEB, Inc.). Amplicons were sequenced on an automated Applied Biosystems 3730 using BigDye Terminator V3.1 cycle sequencing dye terminator chemistry (Perkin-Elmer, Foster City, Calif.). Primer sequences are available upon request. The genomic sequences for Tk/England/50-92/91 (H5N1) and Tk/England/N-28/73 (H5N2) were deposited into GenBank with the accession numbers EU627685, and EU636682 to EU636696.

Datasets

Besides the two archive avian influenza isolates we sequenced, our dataset contains 44,398 influenza gene sequences from Influenza Virus Resource database [61], which were updated in August of 2007. This dataset includes 554 H5N1 AIVs, which have complete or mostly complete genes for all eight genetic segments.

Progenitor gene identification

To identify the potential progenitor genes for each isolate, we applied the newly developed GIN method [59]. Briefly, GIN first measures the genetic distances between viral genes using Complete Composition Vector (CCV) [62], then identifies influenza modules, a cluster of viral genes with small evolutionary distances, using a local optimization program based on thresholds derived from Bayesian analysis [59]. The genes among and within the modules were further analyzed by phylogenetic methods.

Phylogenetic analyses

Multiple sequence alignments of DNA sequences from eight influenza gene segments were performed respectively by the MUSCLE program [63]. The amino acid divergence for specific lineage in the phylogenetic tree was also identified. To reconstruct the tree topology, maximum parsimony, and neighbor-joining method implemented in PAUP* 4.0 Beta [64]. The Maximum Likelihood (ML) tree estimation was evaluated using GARLI version 0.951 [65]. Bayesian inference of phylogeny was performed with BEAST version 1.4.6 [66] with the General Time Reversible (GTR) model estimated by MODELTEST 3.7 [67]. GTR model was run with a gamma distribution modeling rate variation among sites and the proportion of invariable sites for one million iterations (mcmc ngen = 1,000,000). The Tracer version 1.4 was used to estimate the confidence of MCMC analyses from BEAST, and the effective sample size (ESS) must have a minimum of 100 as suggested by the manual. The TreeAnnotator version 1.4.6 was applied to extract the tree with the highest clade credibility. The tree topologies were confirmed for the four methods used. The influenza gene trees (except NP) shown in Figure 2.1 were the Bayesian inference trees from BEAST. The ML tree from GARLI was used for NP gene as shown in Figure 2.1 since ESS from BEAST analysis did not meet our minimum requirement even with 50 million of iterations (mcmc ngen = 50,000,000). Control and log files for all stand-alone programs ran here and other methodological materials are available on request.

Results

Diverse origins for gene segments of H5N1 HPAIVs

Through integrating GIN and phylogenetic tree construction, we identified multiple lineages associated with H5N1 HPAIVs since 1996, each of which contains one or a set

of segments from AIVs isolated before 1996 [59]. The results showed that the HA genes of these H5N1 viruses are phylogenetically close to Tk/England/50-92/91(H5N1) (Tk/E91-like viruses) (Figure 2.1A). However, the NA and six internal gene segments originated from 3 to 7 different lineages: NA has 3 lineages; PB2 has 5; PB1 has 5; PA has 4; NP has 6; M has 7; NS has 2 (Table 2.1; Figure 2.1; Supplementary figure A.1). Most of these pre-96 AIVs were isolated from southeastern or eastern Asia, especially in southern China. Besides these Asian strains, some H5N1 segments are closely related to strains isolated earlier in Europe, such as African starling/England-Q/983/79(H7N1) and Tk/England/N28/73(H5N2) (Figure 2.1; Table 2.1).

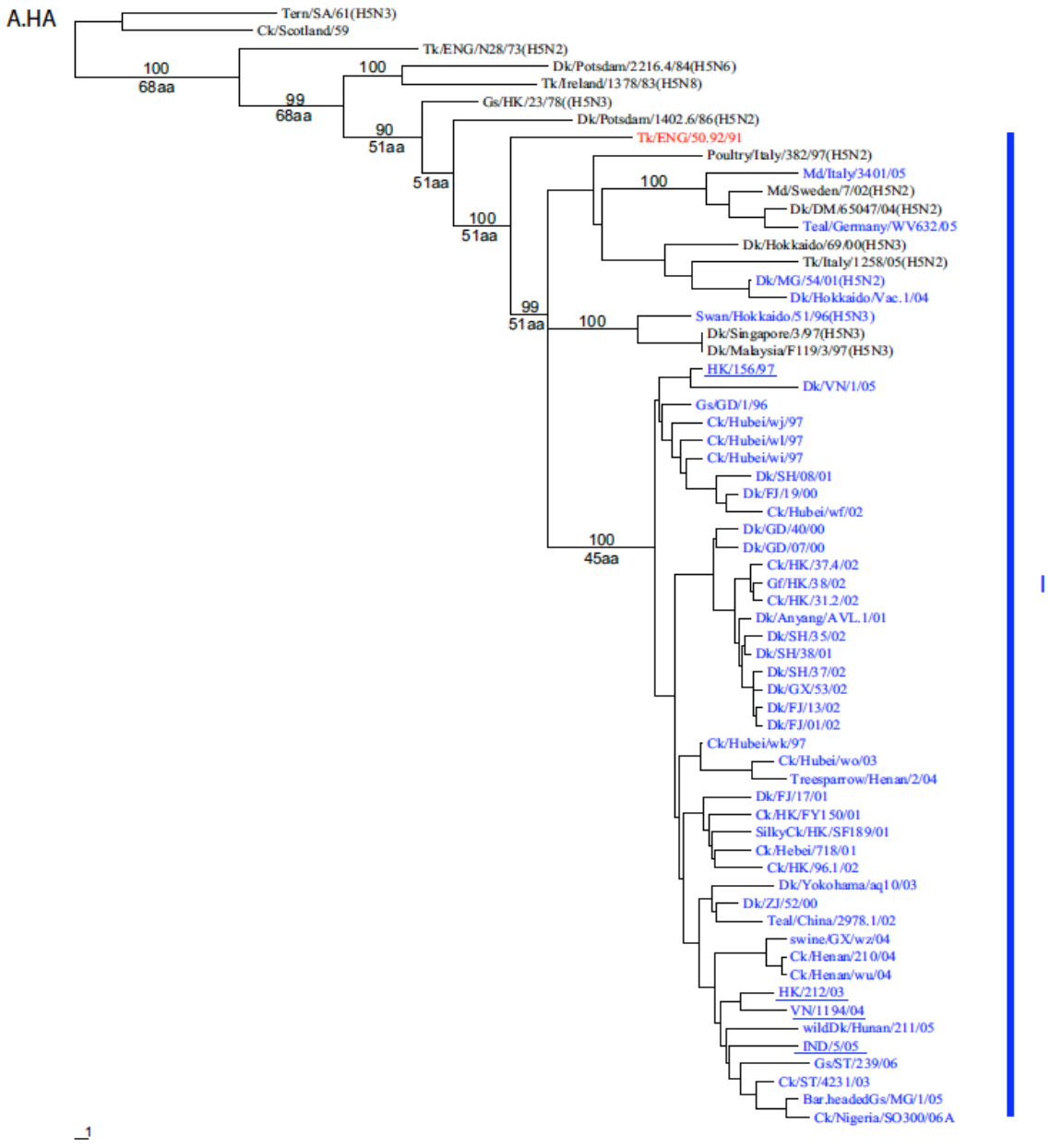


Figure 2.1: Phylogenetic trees of HA(A), NA (B), PB2(C), PB1(D), PA(E), NP(F), MP(G), and NS (H) for H5 HPAIVs and associated progenitor genes.

The trees were rooted by Ck/Scotland/59. The bootstrap analyses were conducted using neighbor joining method implemented in PAUP* based on 1000 replicates. The maximum amino acid divergences were labeled for some interesting lineages. The putative progenitor genes are marked in red, and the human strains were underlined. The lineage number and reassortants are summarized in Tables 1 and 2, respectively.

Figure 2.1 continued

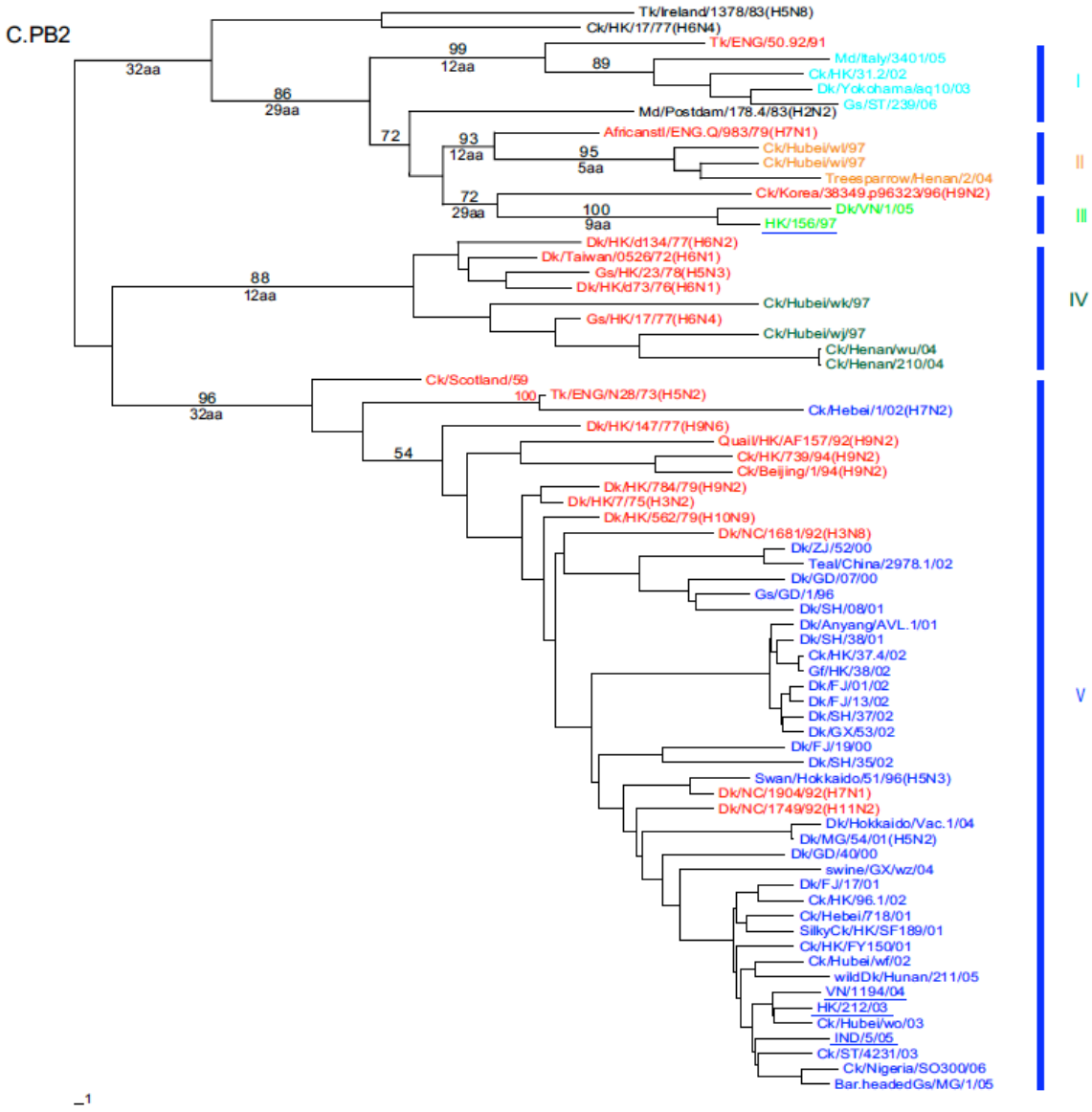


Figure 2.1 continued

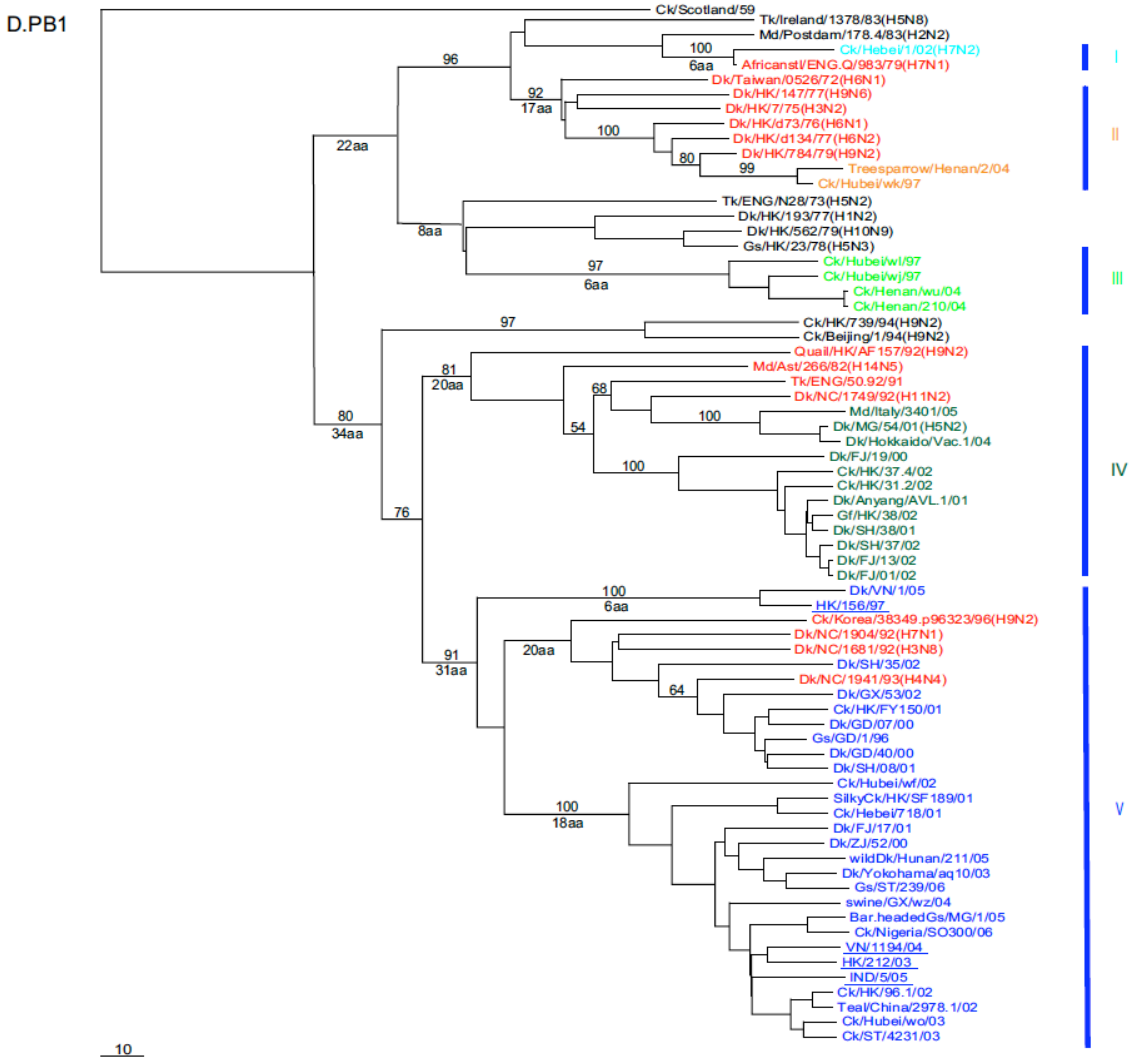


Figure 2.1 continued

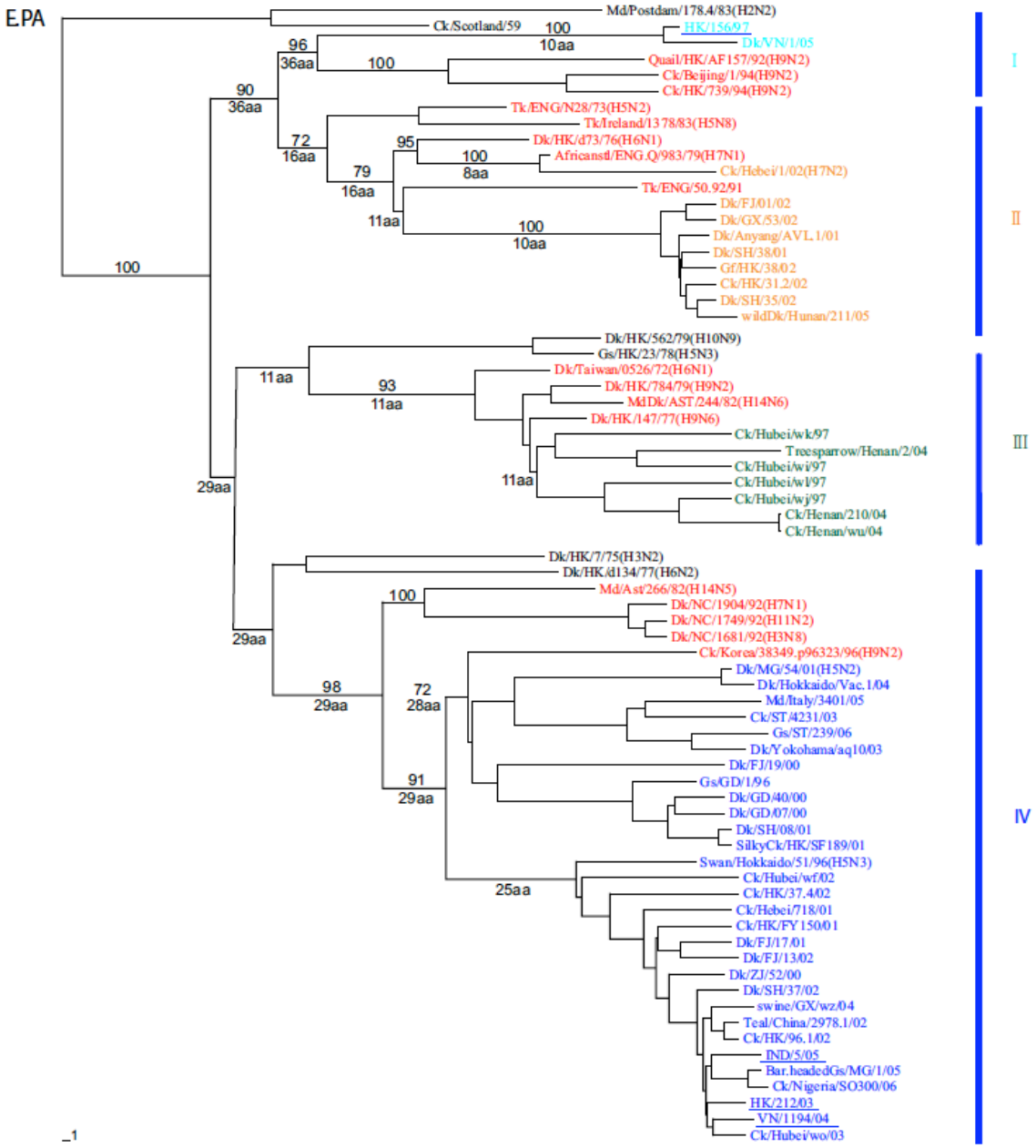


Figure 2.1 continued

Figure 2.1 continued

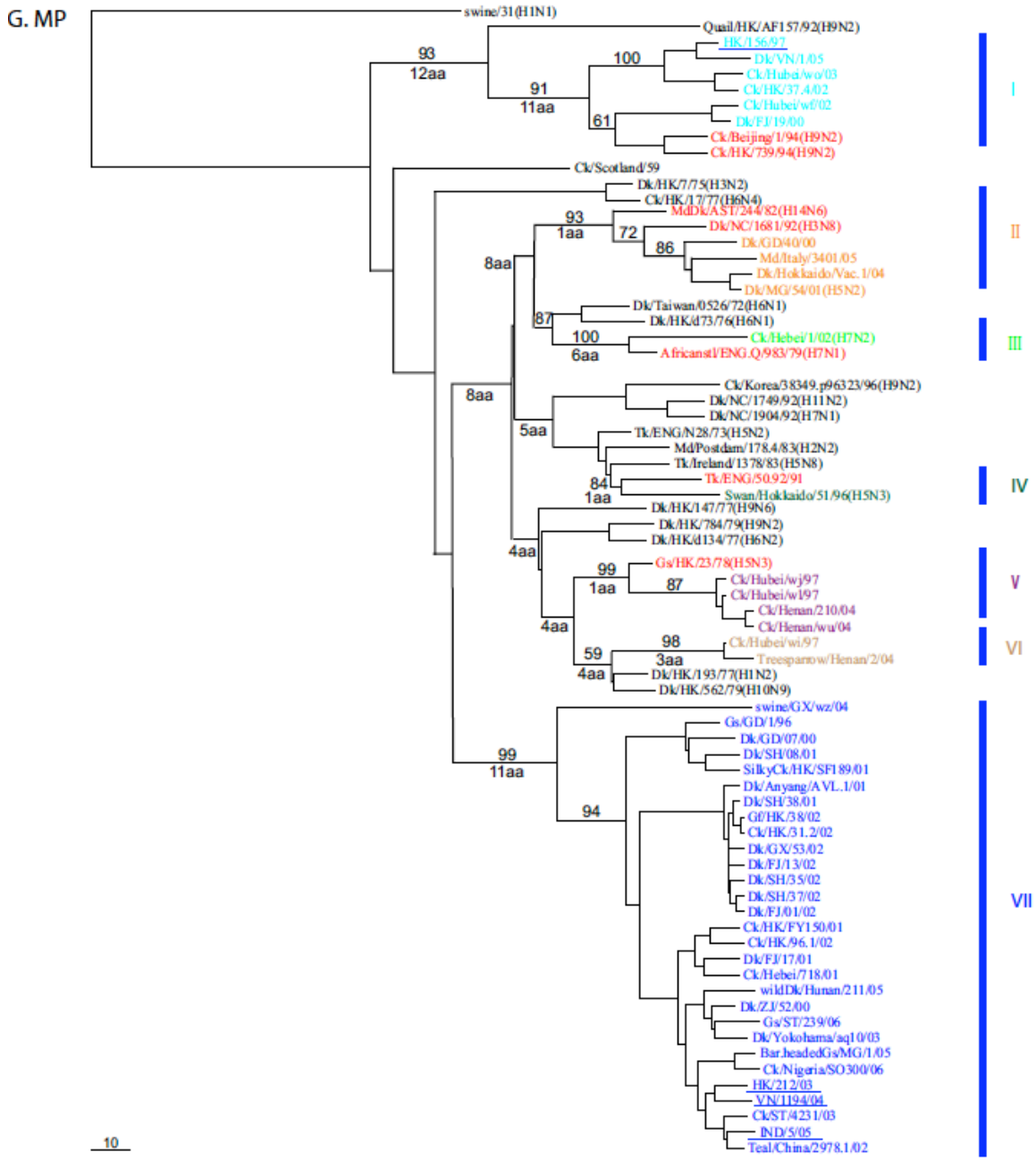


Figure 2.1 continued

Figure 2.1 continued

Table 2.1: Potential progenitor genes identified for H5N1 highly pathogenic AIVs.

Abbreviations: Ast, Astrakhan; Aus, Australia; Ck, chicken; Dk, duck; ENG, England; FJ, Fujian; GD, Guangdong; Gs, goose; GX, Guangxi; HK, Hong Kong; IND, Indonesia; Md, mallard; NC, Nanchang; Stl, starling; Tk, turkey; VN, Vietnam; ZJ, Zhejiang.

Lineages*	Potential progenitor genes for each gene segment							
	PB2	PB1	PA	HA	NP	NA	MP	NS
I	Tk/ENG/50-92/91	African Stl/ENG.Q/ 983/79(H7N1)	Quail/HK/AF157/ 92(H9N2), Ck/ Beijing/1/94(H9N2), Ck/HK/739/ 94(H9N2)	Tk/ENG/50-92/91	Tk/ENG/N28/ 73(H5N2)	African Stl/ENG.Q/ 983/79 (H7N1)	Ck/Beijing/1/ 94(H9N2), Ck/HK/ 739/94(H9N2)	Dk/HK/7/75(H3N2), Dk/HK/193/ 77(H1N2), Dk/NC/ 1681/92(H3N8), Swan/Hokkaido/ 51/96(H5N3), Dk/ Taiwan/0526/ 72(H6N1), Dk/HK/ d134/77(H6N2), Tk/ENG/50-92/91
II	African Stl/ENG.Q/ 983/79(H7N1)	Dk/Taiwan/0526/ 72(H6N1), Dk/HK/ 147/77(H9N6), Dk/ HK/7/75(H3N2), Dk/HK/d73/ 76(H6N1), Dk/HK/ d134/77(H6N2), Dk/ HK/784/79(H9N2)	Tk/ENG/N28/ 73(H5N2), Tk/ Ireland/1378/ 83(H5N8), Dk/HK/ d73/76(H6N1), African Stl/ENG.Q/ 983/79(H7N1), Tk/ ENG/50-92/91		African Stl/ENG.Q/ 983/79(H7N1)	Unknown	MdDk/AST/244/ 82(H14N6), Dk/NC/1681/ 92(H3N8)	African Stl/ENG.Q/ 983/79(H7N1), shearwater/Aus/1/ 72(H6N5), Ck/ Scotland/59, Ck/ Beijing/1/ 94(H9N2), Ck/HK/ 739/94(H9N2), Quail/HK/AF157/ 92(H9N2), Md/Ast/ 266/82(H14N5), Dk/HK/d73/ 76(H6N1), Tk/ Ireland/1378/ 83(H5N8), Gs/HK/ 17/77(H6N4), Dk/ HK/784/79(H9N2), Gs/HK/23/ 78(H5N3), Dk/NC/ 1749/92(H11N2), Tk/ENG/N28/ 73(H5N2), Ck/HK/ 17/77(H6N4), Dk/ HK/147/77(H9N6), Dk/HK/562/79/ H10N9, MdDk/ AST/244/ 82(H14N6), Dk/ NC/1941/ 93(H4N4), Dk/NC/ 1904/92(H7N1)
III	Ck/Korea/ 38349.p96323/ 96(H9N2)	Unknown	Dk/Taiwan/0526/ 72(H6N1), Dk/HK/ 784/79(H9N2), MdDk/AST/244/ 82(H14N6), Dk/HK/ 147/77(H9N6)		Quail/HK/AF157/ 92(H9N2), Ck/ Beijing/1/94(H9N2), Ck/HK/739/ 94(H9N2)	Tk/ENG/50-92/91	African Stl/ENG.Q/ 983/79(H7N1)	
IV	Dk/HK/d134/ 77(H6N2), Dk/ Taiwan/0526/ 72(H6N1), Gs/HK/ 23/78(H5N3), Dk/ HK/d73/76(H6N1), Gs/HK/17/77(H6N4)	Quail/HK/AF157/ 92(H9N2), Md/Ast/ 266/82(H14N5), Tk/ ENG/50-92/91, Dk/ NC/1749/92(H11N2)	Md/Ast/266/ 82(H14N5), Dk/NC/ 1904/92(H7N1), Dk/ NC/1749/ 92(H11N2), Dk/NC/ 1681/92(H3N8), Ck/ Korea/38349.p96323/ 96(H9N2)		Gs/HK/23/78(H5N3)		Tk/ENG/50-92/91	
V	Ck/Scotland/59, Tk/ ENG/N28/73(H5N2), Dk/HK/147/ 77(H9N6), Quail/ HK/AF157/ 92(H9N2), Ck/HK/ 739/94(H9N2), Ck/ Beijing/1/94(H9N2), Dk/HK/784/ 79(H9N2), Dk/HK/7/ 75(H3N2), Dk/HK/ 562/79(H10N9), Dk/ NC/1681/92(H3N8), Dk/NC/1904/ 92(H7N1), Dk/NC/ 1749/92(H11N2)	Ck/Korea/ 38349.p96323/ 96(H9N2), Dk/NC/ 1904/92(H7N1), Dk/ NC/1681/92(H3N8), Dk/NC/1941/ 93(H4N4)			Dk/HK/d134/ 77(H6N2)		Gs/HK/23/78(H5N3)	
VI					Md/Ast/266/ 82(H14N5), MdDk/ AST/244/82(H14N6), Tk/ENG/50-92/91, Ck/Korea/ 38349.p96323/ 96(H9N2), Dk/NC/ 1749/92(H11N2), Dk/NC/1681/ 92(H3N8), Dk/NC/ 1904/92(H7N1)		Unknown	
VII							Unknown	

Combination of progenitor genes

The combinations between these progenitor genes generated at least 21 reassortants from 1996 to 2007, named in this study as H5N1 Progenitor Reassortant (H5N1-PR) 1-21 (Table 2.2). The reassortant nomenclature for the 554 H5N1 AIVs is listed in Supplementary table A.1.

Our results suggest that the surface proteins HA and NA for these reassortants, except H5N1-PR2 and PR10, are likely to be derived from the lineage Tk/E91-like viruses. Instead, the NA gene of PR10 was possibly derived from an African starling/England-Q/983/79(H7N1)-like virus. The progenitor for the NA gene of H5N1-PR2 was not been able to be identified. Nevertheless, the AIVs in eastern Asia such as Gs/HK/23/78(H5N3)-like, Quail/HK/AF157/92(H9N2)-like, Dk/HK/d134/77(H6N2)-like, Dk/NC/1681/92(H3N8)-like, Dk/NC/1749/92(H11N2)-like, and Dk/NC/1904/92(H7N1)-like viruses, constituted a large gene segment reservoir for reassorting as recently put forward [68].

Among the 21 reassortants, H5N1-PR1 contained Gs/Gd/1/96; only H5N1-PR2 and H5N1-PR7 have been identified in human cases. H5N1-PR1 was a reassortment between Tk/E91-like (HA, NA, and NS), Dk/NC/1681/92(H3N8)-like (PB2, PB1, PA, and NP), and an unknown donor for MP (Lineage VII) (Supplementary figure A.1). The only lineage difference between H5N1-PR7 and PR1 is located in NS gene (Supplementary figure A.1) (Table 2.2). NS in PR1 was likely to have been derived from Tk/E91-like viruses while the NS in PR7 has a close phylogenetic distance to Dk/HK/542/79(H10N9)-like and Dk/HK/147/77(H9N6)-like viruses (Figure 2.1 and Supplementary figure A.1).

Frequent reassortments involved some strains isolated 20 to 30 years ago. The NA and NP genes of Ck/Hebei/718/2001(H5N1) (H5N1-PR10) were very close to African starling/England/983/79(H7N1)-like viruses (Figure 2.1) while the other four internal segments had diverse evolutionary origins: its PB2 and PA is close to AIVs isolated in China as well as swan/Hokkaido/51/96 (H5N3). It is worth mentioning that the NP gene of Ck/Hebei/1/02(H7N2) is virtually identical to that of Tk/England/N-28/73 (H5N2) in both nucleotide and amino acid compositions (Figure 2.1F).

Table 2.2: The reassortants from combinations of progenitor genes for H5N1 HPAIVs and some low pathogenic AIVs identified in eastern Asia.

*H5N1-PR denotes H5N1 reassortants from their progenitor genes; O-PR denotes other subtypes of reassortants from H5N1 progenitor genes.

†The reassortment events for 554 H5N1 AIVs with complete genome sets are shown in Supplementary table A.1. The abbreviations are shown in legend of Table 2.1.

‡The lineage numbers were annotated in Figure 2.1 and Supplementary figure A.1.

§H5N1 AIVs isolated from human.

Reassortant*	Selected strain†	Year	No. strains†	Gene segment‡								Chen *et al.* (2006); Guan *et al.* (2002); Li *et al.* (2004)	Chen *et al.* (2004)
				PB2	PB1	PA	HA	NP	NA	MP	NS		
H5N1-PR1	Gs/GD/1/96	1996–2001	12	V	V	IV	I	VI	III	VII	I	Gs/GD	A
H5N1-PR2	HK/156/97§	1997–2005	17	III	V	I	I	VI	II	I	II		
H5N1-PR3	Ck/Hubei/wl/97	1997	1	II	III	III	I	IV	III	V	II		
H5N1-PR4	Ck/Hubei/wj/97	1997–2004	3	IV	III	III	I	V	III	V	II		
H5N1-PR5	Ck/Hubei/wh/97	1997	1	V	II	III	I	V	III	VI	II		
H5N1-PR6	Dk/Yokohama/aq10/03	2003–2006	3	I	V	IV	I	VI	III	VII	II		
H5N1-PR7	Bar headedGs/MG/1/05	1999–2007	478	V	V	IV	I	VI	III	VII	II		
	Ck/HK/NT873.3/01			V	V	IV	I	VI	III	VII	II	A, B, C, E, G, V, W, Y, Z, Z+	B, D, G
	IND/5/05§			V	V	IV	I	VI	III	VII	II		
	VN/1194/04§			V	V	IV	I	VI	III	VII	II		
	ZJ/16/06§			V	V	IV	I	VI	III	VII	II		
H5N1-PR8	Ck/HK/37.4/02	2000–2003	3	V	IV	IV	I	VI	III	I	II	X2	C
H5N1-PR9	Dk/Anyang/AVL.1/01	2001–2004	6	V	IV	II	I	VI	III	VII	II	X0	I, F
H5N1-PR10	Ck/Hebei/718/01	2001	1	V	V	IV	I	II	I	VII	II		
H5N1-PR11	Ck/HK/FY150/01	2001	2	V	V	IV	I	III	III	VII	II	D	
H5N1-PR12	Dk/Hokkaido/Vac.1/04	2001–2004	2	V	IV	IV	I	VI	III	II	II		
H5N1-PR13	Dk/FJ/13/02	2001–2002	3	V	IV	IV	I	VI	III	VII	II	X3	H
H5N1-PR14	Dk/GX/53/02	2002–2005	4	V	V	II	I	VI	III	VII	II		I, F
H5N1-PR15	Ck/HK/31.2/02	2002	1	I	IV	II	I	VI	III	VII	II	X1	
H5N1-PR16	Ck/Hubei/wf/02	2002–2003	2	V	V	IV	I	VI	III	I	II		
H5N1-PR17	Dk/GD/40/00	2000–2005	1	V	V	IV	I	VI	III	II	II		E
H5N1-PR18	Tree sparrow/Henan/4/04	2004	1	V	II	III	I	IV	III	V	II		
H5N1-PR19	swine/GX/wz/04	2004	1	V	V	IV	I	IV	III	VII	II		
H5N1-PR20	Tree sparrow/Henan/2/04	2004	2	II	II	III	I	IV	III	VI	II		
H5N1-PR21	Md/Italy/3401/05	2005	1	I	IV	IV	I	VI	III	II	II		
O-PR1	Swan/Hokkaido/51/96(H5N3)	1996	1	V	V	IV	I	Unknown	NA	IV	I		
O-PR2	Dk/Hokkaido/55/96(H1N1)	1996	1	NA	NA	NA	NA	NA	III	NA	NA		
O-PR3	Dk/MG/54/01(H5N2)	2001	1	V	IV	IV	I	VI	NA	II	I		
O-PR4	Ck/Hebei/1/02(H7N2)	2002	1	V	I	II	NA	I	NA	III	I		

Relationships of H5N1-PRs with reported reassortants

Before our studies, there has been a number of H5N1 genotypes reported in mainland China and Hong Kong which generally identify a reassortant by combining lineages/sub-lineages defined by tree topologies (instead of progenitor genes used in this study) from NA and internal segments [56-58]. Since multiple lineages may be derived from the same progenitor genes, a H5N1-PR may include multiple reassortants reported earlier. For instance, H5N1-PR7 includes genotypes A, B, C, E, G, V, W, Y, Z, and Z+ [58, 69]. H5N1-PR7 also includes the recently reported Qinghai and Fujian-like lineages [50, 69]. As shown in Table 2.2, thirteen H5N1-PRs are first reported in this study.

Emergence of low pathogenic AIVs through progenitor gene combinations

The combinations of the progenitor genes have also resulted in at least four reassortants in eastern Asia: O-PR1, Swan/Hokkaido/51/96(H5N3)-like viruses; O-PR2, Duck/Hokkaido/55/96(H1N1)-like viruses; O-PR3, Duck/Mongolia/54/01(H5N2)-like viruses; O-PR4, Chicken/Hebei/1/02(H7N2)-like viruses (Table 2). Swan/Hokkaido/51/96(H5N3) and Duck/Hokkaido/55/96(H1N1) were identified in Japan in 1996 the same year as Gs/Gd/1/96(H5N1) [70]. The phylogenetic analyses suggest that both HA and NA of Gs/Gd/1/96(H5N1) are likely to be derived from Tk/E91-like viruses (Figure 2.1 and Table 2.2). It is interesting that the HA of Swan/Hokkaido/51/96 (H5N3)-like viruses and the NA of Dk/Hokkaido/55/96(H1N1)-like viruses were likely to have been derived from Tk/E91-like viruses. However, Swan/Hokkaido/51/96(H5N3)-like viruses have internal segments possibly derived from Dk/NC/1904/92(H7N1)-like (PB2, PB1, PA), Tk/England/N28/73(H5N2) (NP)-like, and Tk/E91-like (MP and NS) viruses.

Discussion

On the assumption that the H5N1 HPAIVs have emerged from reassortment events within the AIV gene pool encompassing the Eurasian landmass, this study attempted to identify potential progenitor genes for such viruses identified since 1996. While knowledge of the components of that gene pool remains incomplete, our results demonstrated that the H5N1 viral gene segments have diverse genetic origins, most of which were detected before 1996. Combinations of progenitor genes identified generated at least 21 reassortants, so called H5N1-PR1 to 21, thirteen of which are first reported (Table 2.2 and Supplementary table A.2). The newly developed GIN drew upon publicly available H5N1 virus sequences facilitating a more definitive characterization of the gene pool and nomenclature system, H5N1-PR. Compared with the earlier genotyping system (Guan et al., 2002; Chen et al. 2004), H5N1-PR may provide a more definitive, mutually inclusive research based tool since it focuses on genetic origins instead of reassortment events.

The results suggest that the source of progenitor genes might have a critical impact on host adaptation and/or pathogenesis of these H5N1 viruses. For instance, H5N1-PR1 and PR7 have only a single lineage difference in the NS gene (Supplementary figure A.1D) (Table 2.2). H5N1-PR1, containing Gs/Gd/1/96, was identified in goose, waterfowl, and environmental samples, which generally referred at the time to any samples (faeces or otherwise) found on the poultry floor or cage (without clear host record). However, H5N1-PR7 viruses were identified directly in both waterfowl and land-based birds, such as chickens. Viruses from H5N1-PR7 caused most of the reported outbreaks in both domestic and wild birds and confirmed H5N1 human cases since 2003. In vitro experiments showed that the NS gene enhances virus replication in mammalian cells

[71]. The residue mutation from aspartic acid to glutamic acid at position 92 in the NS1 protein was reported to increase the virulence of avirulent A/Puerto Rico/8/34 (H1N1) in pigs [72]. However, viruses belonging to the NS lineage related to 1997 Hong Kong outbreak have glutamic acid at residue 92 at NS1, and the other H5N1 viruses have aspartic acid at this position (data not shown). Since 2000, there is a similar 5-amino acid deletion at positions 80-84 in the NS1 gene of most H5N1 viruses especially those isolates after the 2003/2004 H5N1 outbreak in eastern Asia [73]. Although viruses without deletions may still circulate in wild birds, e.g. A/mallard/Guangxi/wt/2004(H5N1) and A/slaty-backedgull/Shandong/38/04 (H5N1), it was shown that this deletion can increase the pathogenesis of H5N1 viruses in chickens [74].

The findings indicate that the PR2 reassortant would have been involved with the 18 cases recorded in the H5N1 outbreak in Hong Kong in 1997 which included 6 fatalities [48]. This reassortant disappeared after slaughtering 1.5 million chickens there and stopping live poultry trade for 7 weeks [48]. H5N1 97-like viruses have been recorded once since 1997 being from egg shell washes taken from one goose and two duck eggs imported from Vietnam into China in 2005 [75]. The reason for this is unclear. It is not impossible that H5N1-PR2 reappearing in 2005 might have been detected as a consequence of HK97-like virus inactivated vaccine usage or lab contamination. Nevertheless it raises important issues on H5N1 ecology warranting intensive virus surveillance in the region. H5N1-PR2 has 4 different progenitor genes from H5N1-PR1, which includes Gs/Gd/1/96(H5N1). Thus, the 1997 Hong Kong outbreaks may have originated from a different reassortment event from that gave rise to Gs/Gd/1/96 (H5N1).

Besides these reassortants, several other H5N1 HPAIVs, such as Ck/Hubei/wi/97, Ck/Jilin/hg/02, WildDk/GD/314/04, and Ck/Hubei/wk/97, have different combinations of the progenitor genes identified in this study (Supplementary table A.1). However, some genotypes did not form a well supported lineage together with progenitor genes or lacked complete genomic datasets and thus are not included in Table 2. For instance, the MP gene of WildDk/GD/314/04 did not cluster with any other viruses to form a well supported lineage with any progenitor genes. Therefore, the 21 H5N1 reassortants identified in this study are still incomplete; the emergence of new reassortants apparently continues. By virtue of providing further genetic background to the origins of H5N1 HPAIVs, this study's findings could be useful toward developing an influenza prevention and control strategy. Such strategy must be based on long term systematic AIV surveillance, quick provision of gene sequences and isolates as appropriate and structured international coordination. Genetic data available for pandemic viruses of the 20th Century indicate that they are unlikely to have been highly pathogenic for chickens and other types of bird (www.oie.int). It remains to be seen whether there are other H5N1 viruses that have been stored in Asia or elsewhere, as yet ungenotyped, that could shed light on the functional range of progenitor genes.

Abbreviation

HPAIVs, Highly Pathogenic Avian Influenza Viruses; AIVs, Avian Influenza Viruses; HA, Hemagglutinin; NA, Neuraminidase; NP, Nucleoprotein; PR, Progenitor Reassortant; CCV, Complete Composition Vector; ML, Maximum Likelihood; GTR, General Time Reversible; ESS, effective sample size; GIN, Genotype In Network.

Acknowledgements

This work was supervised by Dr. Xiufeng (Henry) Wan. It was supported by a Miami University CFR grant and NSF Award BCS-0717688 to Xiufeng (Henry) Wan.

CHAPTER 3

EVOLUTION OF H5N1 HIGHLY PATHOGENIC AVIAN INFLUENZA VIRUSES IN VIETNAM BETWEEN 2001 AND 2007 [76]

Abstract

Phylogenetic analyses of eight genetic segments of H5N1 highly pathogenic avian influenza viruses (HPAIVs) in Vietnam between 2001 and 2007 showed that the viruses were introduced into Vietnam multiple times, and experienced multiple times of reassortments. Hemagglutinin (HA) and neuraminidase (NA) were introduced into Vietnam six times independently from precursor viruses in China, and reassorted with other six internal segments to give rise to multiple genotypes circulating in Vietnam.

Introduction

H5N1 HPAIVs caused tremendous economic loss and threatened public health in Vietnam since its first introduction in 2001 [77]. H5N1 HPAIVs have been continuously isolated and sequenced by National Center for Veterinary Diagnostics (NCVD) in Hanoi since 2001 till our studies in 2007. For our studies, we included over 300 complete or nearly complete H5N1 HPAIVs, and we aimed to understand the origins and evolution of these H5N1 viruses for virus control and prevention. Our studies indicated that nine geneotypes were present in Vietnam via viral introduction from China and further reassortments, and five genotypes were circulating in 2007 [76].

Methods

Datasets and phylogenetic analyses.

We have used 333 AIVs in Vietnam from 2001 to 2007 and all datasets can be retrieved from our paper [76]. Gene In Network (GIN) method [59] was first applied to each of the eight genetic segments to detect the clusters of H5N1 viruses. GIN measures the evolutionary distance between genes using the Complete Composition Vector (CCV) approach [59, 62, 78]. Representative strains from each cluster were selected for extensive phylogenetic analyses.

Phylogenetic analyses were performed by Maximum Parsimony (MP) and Neighbor-Joining (NJ) methods using PAUP* 4.0 Beta [64]. Maximum Likelihood (ML) tree estimation was evaluated using GARLI version 0.951 [65]. Bayesian trees were estimated using MrBayes version 3.1.2 [79] from 1 million generations, sampling every 100 generations, with the default heating parameter, in two runs. The consensus trees were calculated using allcompat option from the final 10,001 trees of each run. Tree topologies were confirmed between each of these three methods. Bootstrapping support for tree topologies were performed using NJ methods implemented in PAUP* 4.0 Beta with 1,000 replicates. When Bayesian trees were estimated, the posterior probability for each split was generated using the MrBayes sumt option with a 25% burnin. These posterior probabilities were used as an alternative measure of clade support. The nucleotide substitution models for ML and NJ methods were selected using MODELTEST 3.7 [67].

Results and discussion

Emergence of H5 HA genes of AIVs in Vietnam

Our phylogenetic analyses of HA showed six well supported clades of Vietnam H5N1 HPAIVs with both posterior probability and bootstrap values labeled in the critical branches (Figure 3.1). All six clades grouped together with certain viruses isolated from

China, thus the clades were labeled by using their precursor viruses in China. The WHO nomenclatures [80] were also mapped to our phylogenetic tree (Figure 3.1). Including Dk/Vietnam/342/01 (H5N2), there are totally seven clades of HA, indicating seven independent introductions of HA to Vietnam.

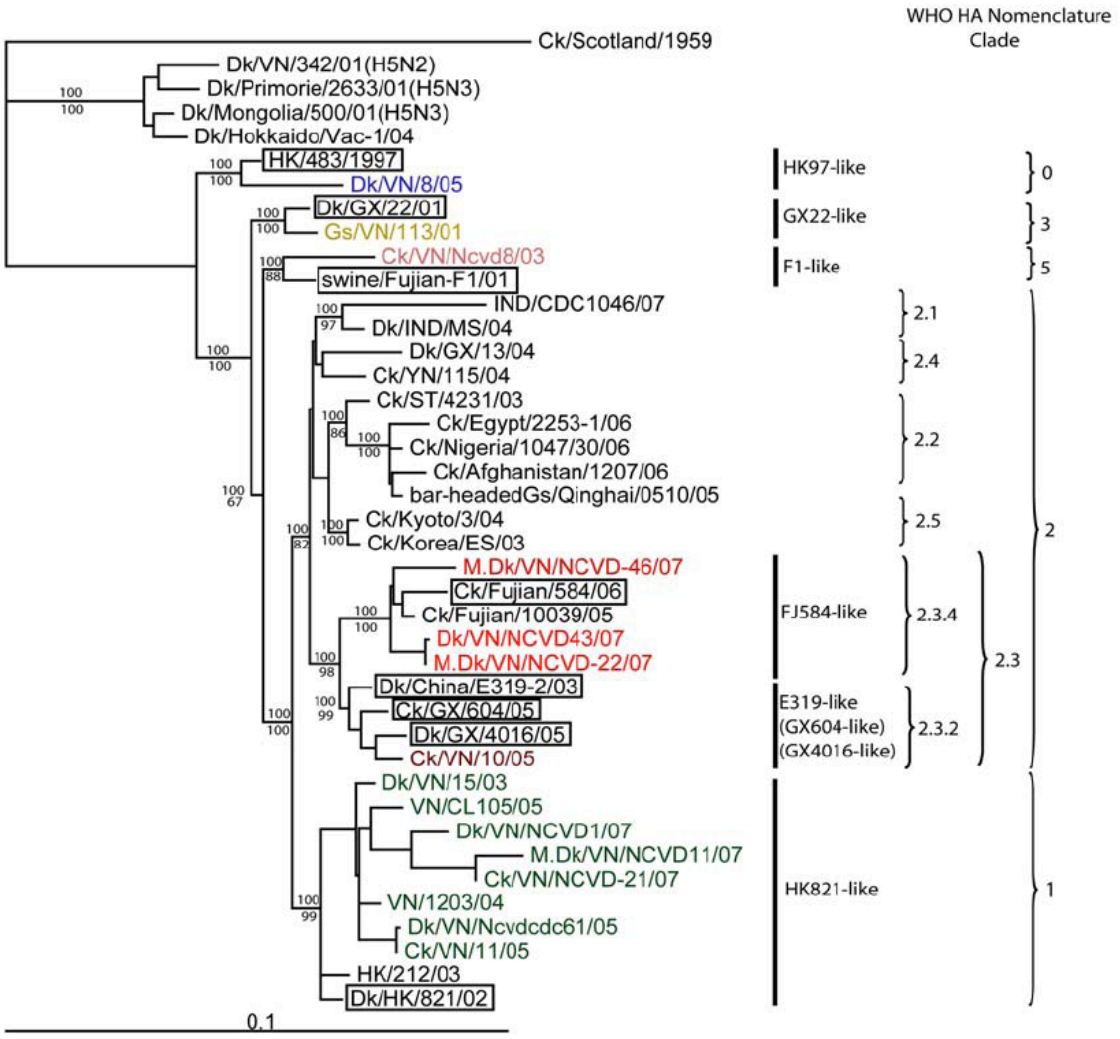


Figure 3.1: Phylogenetic tree of subtype H5 HA genes from avian influenza viruses.
The phylogenetic tree was constructed by Maximum Likelihood using GARLI version 0.951 [34] by selecting GTR+I+G model from Modeltest 3.7 [37]. Posterior probabilities and bootstrap values were given above and below branches, respectively. The phylogenetic tree was rooted by Ck/Scotland/59. Each avian influenza gene precursor was defined with the shortest phylogenetic distance to the Vietnam lineages: HK97-, HK821-, E319-, FJ584-, GX22-, and F1-, GX604-, and GX4016-like. The viruses isolated from Vietnam are marked in colors other than black. The predicted precursor viruses are shown in boxes. A detailed tree showing the HK821-like lineage can be viewed in Supplementary figure B.2.

Phylogenetic Analyses of NA and Internal Genes Revealed an Abundant Genome Segment Pool in Vietnam

Phylogenetic analyses were performed to NA and other six internal genetic segments (Figure 3.2 and Supplementary Figure B.1). NA showed six independent well supported lineages (Figure 3.2). PB2, PB1, NP, MP, and NS genes showed four distinct lineages,

and PA and NA genes showed five distinct lineages (Supplementary Figure B.1). These analyses indicate the presence of multiple introductions and reassortment events to give rise to nine genotypes of H5N1 HPAIVs in Vietnam between 2001 and 2007 [76]. By integrating our genetic analyses with the temporal and spatial distribution of the viruses, we could infer that new genotypes were first generated in northern Vietnam, and the viruses might spread from north to south [76]. Virus control and prevention should be focused on preventing the introduction and reassortment of AIVs.

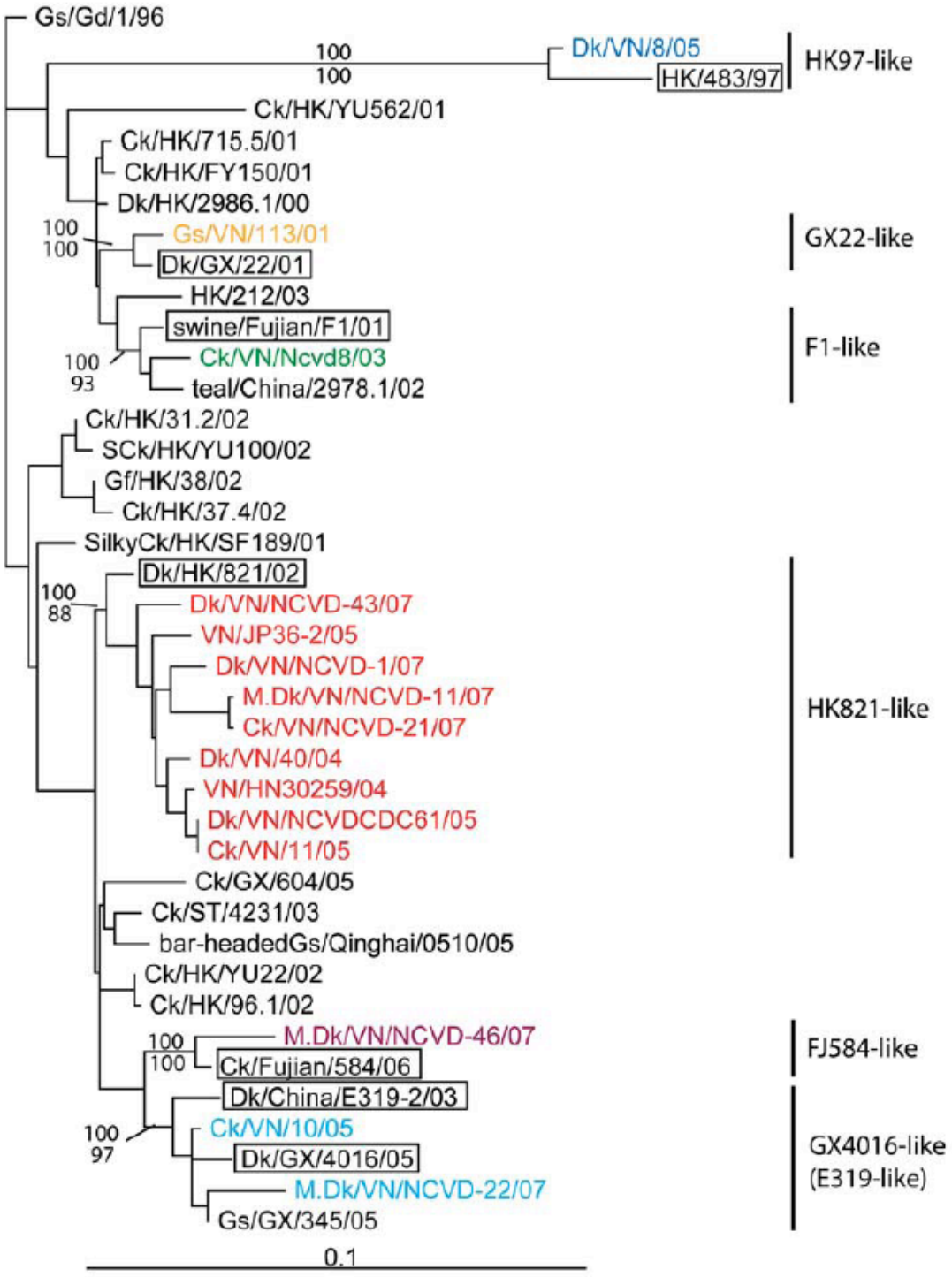


Figure 3.2: Phylogenetic tree of the NA gene of HPAIV H5N1 viruses.

The phylogenetic tree was constructed by Maximum Likelihood using GARLI version 0.951 [34] by selecting GTR+G model from Modeltest 3.7 [37]. Posterior probabilities and bootstrap values were given above and below branches, respectively. The tree was rooted by Gs/GD/1/96. The predicted precursor viruses were shown in boxes.

Abbreviation

HPAIVs, Highly Pathogenic Avian Influenza Viruses; AIVs, Avian Influenza Viruses; HA, Hemagglutinin; NA, Neuraminidase; NCVD, National Center for Veterinary Diagnostics; GIN, Genotype In Network; CCV, Complete Composition Vector; ML, Maximum Likelihood; MP, Maximum Parsimony; NJ, Neighbor-Joining.

Acknowledgements

This work was performed under the supervision of Dr. Xiufeng (Henry) Wan.

CHAPTER 4

THE EVOLUTIONARY HISTORY OF THE CATENIN GENE FAMILY DURING METAZOAN EVOLUTION [17]

Abstract

Catenin is a gene family composed of three subfamilies; p120, beta and alpha. Beta and p120 are homologous subfamilies based on sequence and structural comparisons, and are members of the armadillo repeat protein superfamily. Alpha does not appear to be homologous to either beta or p120 based on the lack of sequence and structural similarity, and the alpha subfamily belongs to the vinculin superfamily. Catenins link the transmembrane protein cadherin to the cytoskeleton and thus function in cell-cell adhesion. To date, only the beta subfamily has been evolutionarily analyzed and experimentally studied for its functions in signaling pathways, development and human diseases such as cancer. We present a detailed evolutionary study of the whole catenin family to provide a better understanding of how this family has evolved in metazoans, and by extension, the evolution of cell-cell adhesion. All three catenin subfamilies have been detected in metazoans used in the present study by searching public databases and applying species-specific BLAST searches. Two monophyletic clades are formed between beta and p120 subfamilies using Bayesian phylogenetic inference. Phylogenetic analyses also reveal an array of duplication events throughout metazoan history. Furthermore, numerous annotation issues for the catenin family have been detected by our computational analyses. Delta2/ARVCF catenin in the p120 subfamily, beta catenin

in the beta subfamily, and alpha2 catenin in the alpha subfamily are present in all metazoans analyzed. This implies that the last common ancestor of metazoans had these three catenin subfamilies. However, not all members within each subfamily were detected in all metazoan species. Each subfamily has undergone duplications at different levels (species-specific, subphylum-specific or phylum-specific) and to different extents (in the case of the number of homologs). Extensive annotation problems have been resolved in each of the three catenin subfamilies. This resolution provides a more coherent description of catenin evolution.

Introduction

Catenin (derived from "catena", "chain" in Latin) is a gene family that links the transmembrane protein cadherin to the cytoskeleton and functions in cell-cell adhesion [81]. The catenin family is composed of three subfamilies; p120 subfamily, beta subfamily and alpha subfamily. The p120 subfamily includes seven members, which are p120 (also named delta1 catenin), Armadillo Repeat protein deleted in Velo-Cardio-Facial syndrome (ARVCF), delta2 catenin (also named NPRAP/Neurojungin), plakophilin (pkp) 4 (also named p0071), pkp1, pkp2, and pkp3 [82]. The beta subfamily includes gamma catenin (also named plakoglobin) in addition to beta catenin [83]. The alpha subfamily includes alpha1 catenin (also named alpha-E-catenin), alpha2 catenin (also named alpha-N-catenin) and alpha3 catenin (also named alpha-T-catenin) [84].

The biomolecular functions and evolutionary history for the catenin family have not been well studied except for beta catenin [83]. Beta catenin functions in both cadherin-associated cell adhesion (adherens junction) and Wnt signaling pathways. This subfamily participates in development and human diseases such as cancer [81, 83]. Previous

evolutionary analyses demonstrate that gamma catenin duplicated from beta catenin in vertebrates and there have been two separate beta catenin duplications specific to *Caenorhabditis elegans* [83]. Gamma (also named plakoglobin) and the four plakophilins are all components of the desmosome [85]. Previous studies have also shown a relationship between p120 and human cancer [82], while limited evolutionary information has been revealed for the p120 subfamily [86].The detailed functions and histories of the p120 and alpha subfamilies, however, have not yet been extensively studied. Thus evolutionary studies for the whole catenin family seem appropriate given the sufficient amount of metazoan sequence data currently available and the interest of the biomedical community in the catenin family. Additionally, the availability of the genomic sequence from the premetazoan unicellular choanoflagellate *Monosiga brevicollis* [87] enables us to make inferences about the catenin family before the emergence of metazoans.

In the present study, we analyzed the gene presence/absence for members of the catenin family from fully sequenced representative metazoans by searching public databases and applying species-specific BLAST searches. All three subfamilies (but not all members of each subfamily) are present in all metazoan species analyzed here. This implies that the last common ancestor of metazoans had all three catenin subfamilies, but members of catenin subfamilies have duplicated at different points in evolutionary history. We applied Bayesian phylogenetic analysis for the p120 and beta subfamilies together, and each of the three catenin subfamilies separately. The results demonstrate that p120 and beta catenins form monophyletic groups and that catenins have undergone

multiple duplications at different levels (species-specific, subphylum-specific or phylum-specific) and to different extents (in the case of the number of paralogs).

Our bioinformatics and evolutionary analyses have also helped to resolve some existing annotation issues associated with catenin members in both vertebrates and non-vertebrates. The annotation issues include wrong annotations, confusing annotations (different names related or not related to the true name), and hypothetical annotations. Confusing or hypothetical annotations for unknown genes or gene families exist because they have not been functionally well characterized. Wrong annotations occur frequently in highly similar or highly divergent homologs, since it is hard to resolve annotation by sequence information alone [88]. Thus, the application of evolutionary analyses to the annotation of new sequences in large gene families can hold considerable value.

Methods

Gene presence/absence for the catenin family

A list of species with complete genome sequences was chosen to represent the species tree for metazoan evolution together with the premetazoan unicellular choanoflagellate *M. brevicollis* serving as the outgroup (Figure 1). The presence/absence of catenin was determined in all the selected species by searching gene names in public databases and/or via species-specific BLAST searches. For all BLAST analyses conducted in the present study, hits having E-values less than 0.05 and sequence identity greater than 15% were considered significant. Detailed BLAST searches against genomic databases were performed for non-vertebrates only in order to determine the history of catenins before the origin of vertebrates. Species-specific duplications in vertebrates have occurred for catenin subfamilies but are not the focus of our current study [89].

Datasets

The protein sequences were extracted for evolutionary analyses and sequence identifiers can be found in the additional files. All sequence identifiers are NCBI protein sequence GI numbers with only two exceptions. The first exception is the pkp1 gene from frog. It was downloaded from the Joint Genome Institute (JGI) *Xenopus tropicalis* assembly v4.1 [31], and it can be queried using the protein ID '156321'. The second exception is the gamma catenin from finch. The sequence is derived from mRNA sequence in the NCBI nucleotide database (nucleotide identifier: 224086507) and translated by the Translate tool on the ExPASy server [77].

Phylogenetic analyses and tertiary structures

Sequence alignments were performed by ClustalW [90]. Phylogenetic analysis was carried out using MrBayes 3.1.2 Unix version [79]. The MPI version of MrBayes was run in parallel on eight nodes with MPICH2 installed [91]. Bayesian trees with posterior probabilities were constructed with mixed amino acid models, a gamma distribution for rate variation among sites, and a proportion of invariable sites. MrBayes was executed with two runs (four chains for each run), one million generations of Markov Chain Monte Carlo (MCMC) analyses, with 1000 as the sample frequency. The number of MCMC generations guaranteed the convergence of two runs by having the standard deviation of split frequencies less than 0.05. The posterior probability of each split was estimated by sumt with 250 trees discarded as burnin based on the plot of 'generation vs. log probability'. Trees with branch lengths and posterior probabilities are shown in Supplementary figure C.1. Parameters were summarized by sump with 250 burnin, and values for the Potential Scale Reduction Factor (PSRF) were all close to 1.0 for all

parameters. The tertiary structures for representative genes were generated by using PyMOL [76].

Annotation validation

Sequence similarity comparisons were conducted by BLAST [92]. Pairwise distance comparisons were conducted using MEGA4 [93] with the Dayhoff model [94] and the JTT model [33]. Statistical significance of differences between two groups of pairwise distances was evaluated by Wilcoxon rank and Kolmogorov-Smirnov tests implemented in R [4]. Bayesian trees with posterior probabilities were constructed to determine evolutionary relationships. Functional divergence among homologs was inferred using the DIVERGE software package [95].

Tissue specific gene expression

We accessed the BioGPS database to identify divergent patterns in gene expression and attempt to correlate potential functional differences between catenin members. BioGPS (an online resource containing gene expression data based on Affymetrix microarrays) was used to visualize and compare gene expression patterns for catenin family members and catenin-related genes in human and mouse cells [96].

Results

Origins of the catenin subfamily members during metazoan evolution

Figure 4.1 shows a cladogram for metazoan evolution [97] with the premetazoan choanoflagellate *M. brevicollis* as the outgroup used for our study. We selected species in the phyla of Vertebrata, Urochordata, Arthropoda, Nematoda, and Cnidaria to represent metazoans. We used Mammalia, Aves, Amphibia and Ray-finned fish to represent classes for Vertebrata; *Homo sapiens* (human) for Mammalia, *Gallus gallus* (chicken) or *Taeniopygia guttata* (finch) for Aves, *Xenopus laevis* or *Xenopus tropicalis* (frog) for

Amphibia, and *Danio rerio* (fish) for Ray-finned fish. For Aves and Amphibia, chicken/finch and *X. laevis*/*X. tropicalis* were used depending on the sequence availability, sequence length and quality. We used *Ciona intestinalis* (sea squirt) for Urochordata, *Drosophila* (fruit fly) and/or other insects for Arthropoda, *C. elegans* and/or others for Nematoda, *Nematostella vectensis* (sea anemone) and/or *Hydra magnipapillata* for Cnidaria. Vertebrata and Urochordata together formed the Chordata phylum, and Bilateria and Cnidaria containing Radiata formed the Metazoa clade. Based on this species tree of metazoan evolution, we determined presence/absence of catenin family members.

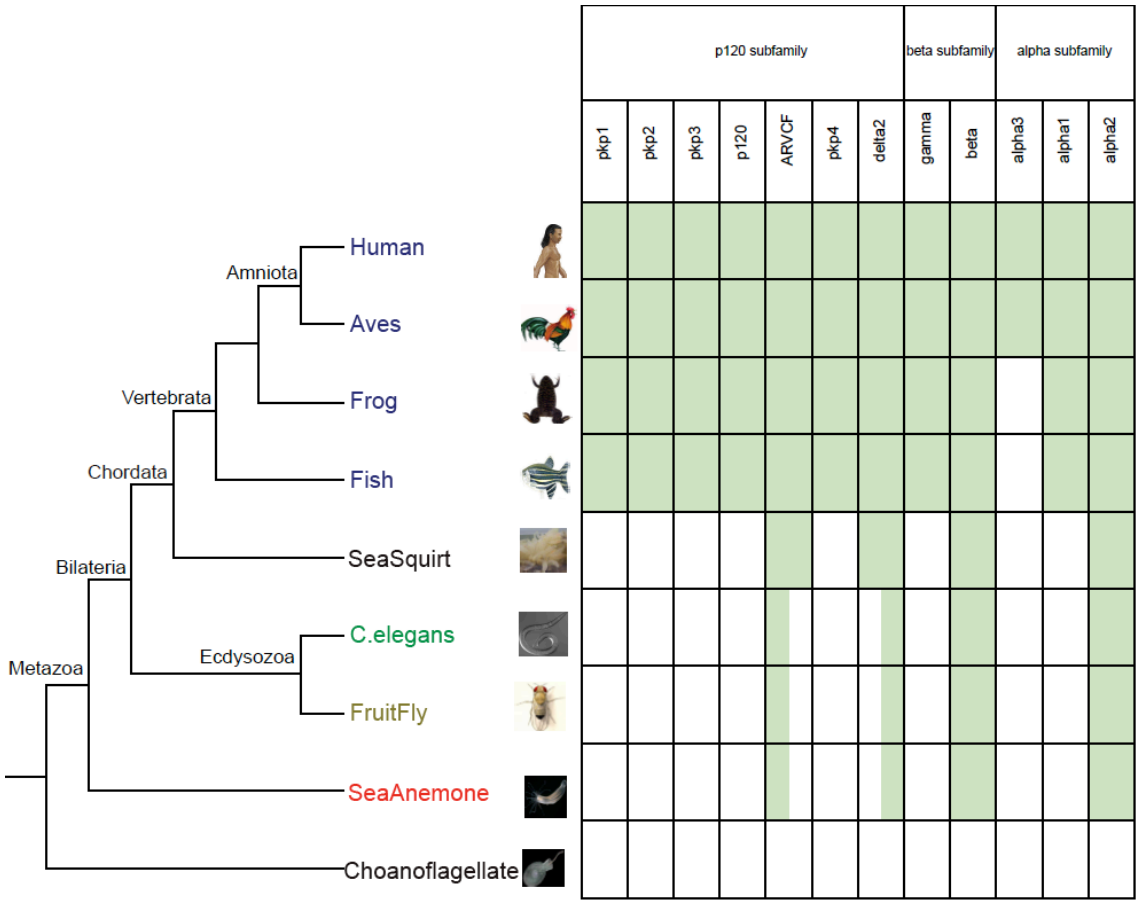


Figure 4.1: The presence/absence of catenin family members in species during metazoan evolution.

The cladogram of metazoans on the left is based on metazoan evolution with the premetazoan choanoflagellate as the outgroup. Species have been selected in the phyla of Vertebrata, Urochordata, Arthropoda, Nematoda, and Cnidaria to represent metazoans. Vertebrata and Urochordata together formed the Chordata phylum, Arthropoda and Nematoda formed the Ecdysozoa clade, Chordata and Ecdysozoa formed the Bilateria clade, and Bilateria and Cnidaria containing Radiata formed the Metazoa clade. We applied different colors for species in different phyla, such as blue for all vertebrates, black for the urochordate sea squirt and the premetazoan choanoflagellate, green for the nematode *C. elegans*, brown for the arthropod fruit fly, and red for the cnidarian sea anemone. Same color patterns were applied for species from different phyla in Figure 4.3, Figure 4.4, and Supplementary figure C.1. Based on the species tree, we determined presence/absence of catenin family members. The table on the right side of the species tree includes all catenin genes, and each shaded green box indicates the presence of the specific gene in the particular species shown in the tree, while the white box indicates the absence of the specific gene. The partially-shaded green boxes indicate unresolved presence/absence of the delta2 and ARVCF members. These species had either a delta2 or an ARVCF but the current analyses cannot resolve the relationship.

In Figure 4.1, green shaded boxes in the table indicate the presence of catenin family members detected in the representative organisms from the species tree. The distribution demonstrates that all three subfamilies have been detected in all the selected metazoan

species but not in the premetazoan *M. brevicollis*. Furthermore, we performed presence/absence analyses to other non-metazoans including the amoeba *Dictyostelium discoideum*, the protist *Cryptosporidium parvum*, three fungi (*Aspergillus*, *Saccharomyces*, and *Schizosaccharomycetes*), three unicellular algae (*Ostreococcus*, *Chlamydomonas*, and *Thalassiosira*), and the plant *Arabidopsis.* None of the catenin members have been detected in the above non-metazoans. We infer that these three catenin subfamilies were therefore present in the last common ancestor of metazoans, but none were present in non-metazoans. All members of the p120 subfamily except ARVCF and delta2 are present in only vertebrates. This scenario also holds for gamma catenin from the beta subfamily and alpha1 catenin from the alpha subfamily. The annotation of the catenin members in non-vertebrates is based on extensive analyses and discussed in the annotation section.

We further extended presence/absence analyses to other non-catenin gene families involved in desmosome formation. These included both desmosomal cadherins desmogleins and desmocollins, as well as desmoplakins. All of these genes have been detected only in vertebrates but not detected in non-vertebrates.

Evolution of an ancient duplication in the catenin family

Figure 4.2 shows that the p120 subfamily forms a monophyletic clade supported by a posterior probability (PP) equal to 0.8 and the beta subfamily forms a separate monophyletic clade (PP = 1.0). The alpha subfamily does not appear to be homologous with p120/beta according to sequence and structural analyses. Representative tertiary structures from the three catenin subfamilies were extracted from the Protein Data Bank [98]: PDB ID accessions 3L6X, 2Z6G, and 1L7C for the p120, beta and alpha

subfamilies, respectively. Figure 4.2 shows cartoon representations of the tertiary structures for the alpha and the p120/beta subfamilies. To the eye, these structures clearly lack analogous folds. Specifically, the p120/beta subfamilies contain armadillo domain (ARM) repeats, while the alpha subfamily contains vinculin homolog domains. An attempt to align the three subfamilies failed to identity conserved anchors that would have allowed us to align homologous regions. Thus, our structural and sequence analyses suggest the alpha subfamily is not evolutionarily related to the p120/beta subfamilies, and the catenin family has undergone an ancient duplication resulting in the p120 and beta subfamilies.

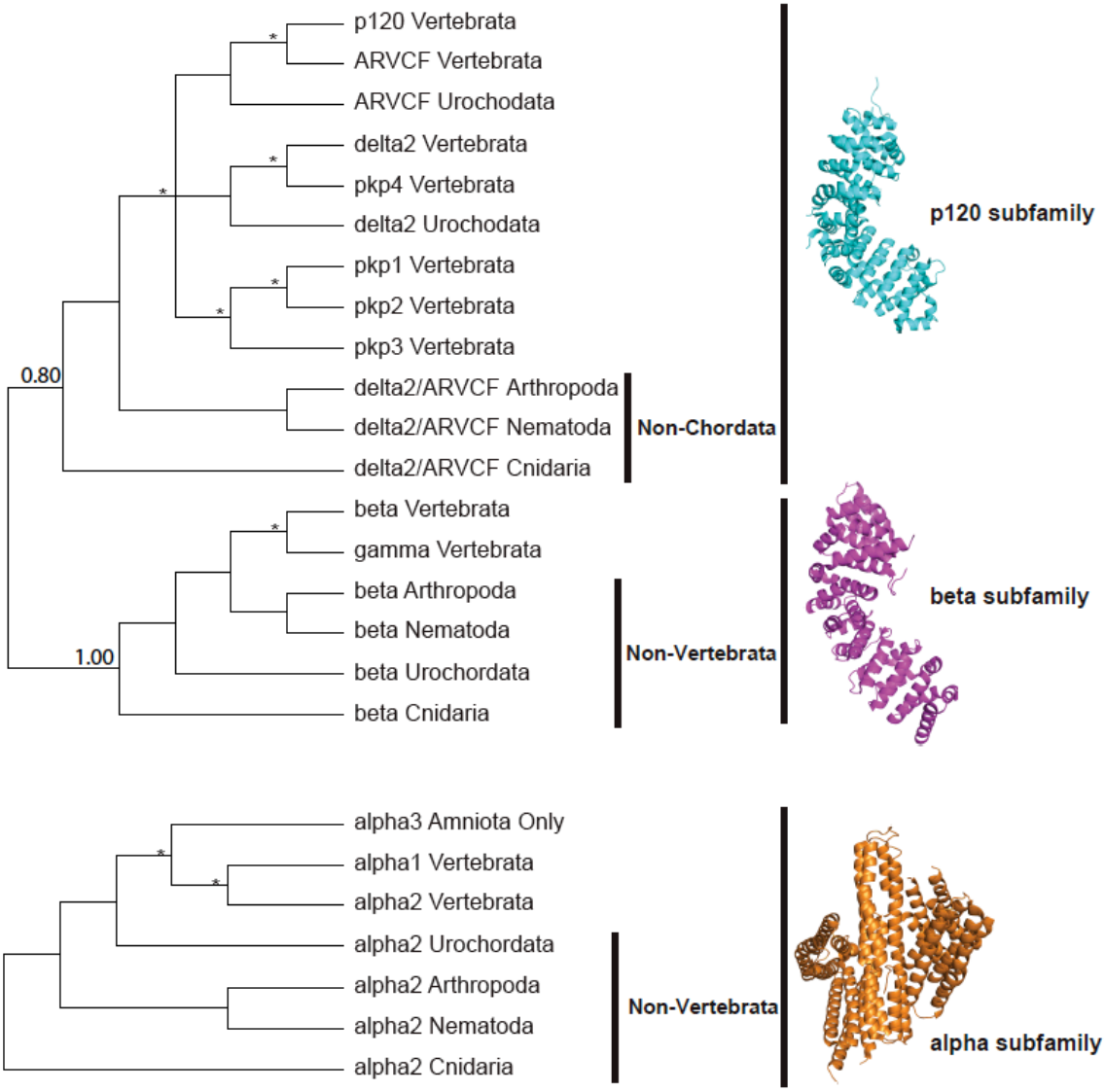


Figure 4.2: The evolution of the catenin family.

The cladogram of each catenin subfamily is based on the individual subfamily topology with sea anemone from the Cnidaria phylum as the outgroup. The tree was condensed to the higher level (the phylum level), including the phyla of Vertebrata, Urochordata, Arthropoda, Nematoda, and Cnidaria. Non-Chordata and Non-Vertebrata were labelled additionally. The taxon names in the tree are a combination of the gene name and the phylum name. The p120 subfamily and the beta subfamily are homologous, thus were shown together in the phylogeny. Asterisks indicate duplication events. The tertiary structures were presented along the phylogeny, with PDB ID accessions 3L6X, 2Z6G, and 1L7C for the p120, beta and alpha subfamilies, respectively.

Origins and duplications of p120 subfamily members

Figure 4.3 shows the Bayesian phylogeny for p120 subfamily members with delta2/ARVCF catenin from sea anemone serving as the outgroup. The complete tree

with branch lengths and NCBI sequence identifiers is shown in Supplementary figure C.1-A. Figure 4.3 shows that p120 and ARVCF in vertebrates form a clade (PP = 1.0) with ARVCF from sea squirt as the outgroup (PP = 1.0). This figure also shows that delta2 catenin and pkp4 in vertebrates form a clade (PP = 1.0) with delta2 from sea squirt as the outgroup (PP = 1.0). Figure 4.3 shows that pkp1, pkp2 and pkp3 in vertebrates form a clade (PP = 1.0) within which pkp1 and pkp2 form a monophyletic clade (PP = 1.0). All of the above genes in chordates form a monophyletic clade (PP = 1.0) with all non-chordate delta2/ARVCF positioned outside the clade. In total, these results suggest that: p120 and ARVCF in vertebrates share a common ancestor with ARVCF in the urochordate; delta2 and pkp4 in vertebrates share a common ancestor with delta2 catenin in the urochordate; pkp1, pkp2 and pkp3 in vertebrates share a common ancestor, and all of the above genes share a common ancestor with delta2/ARVCF catenin in non-chordates.

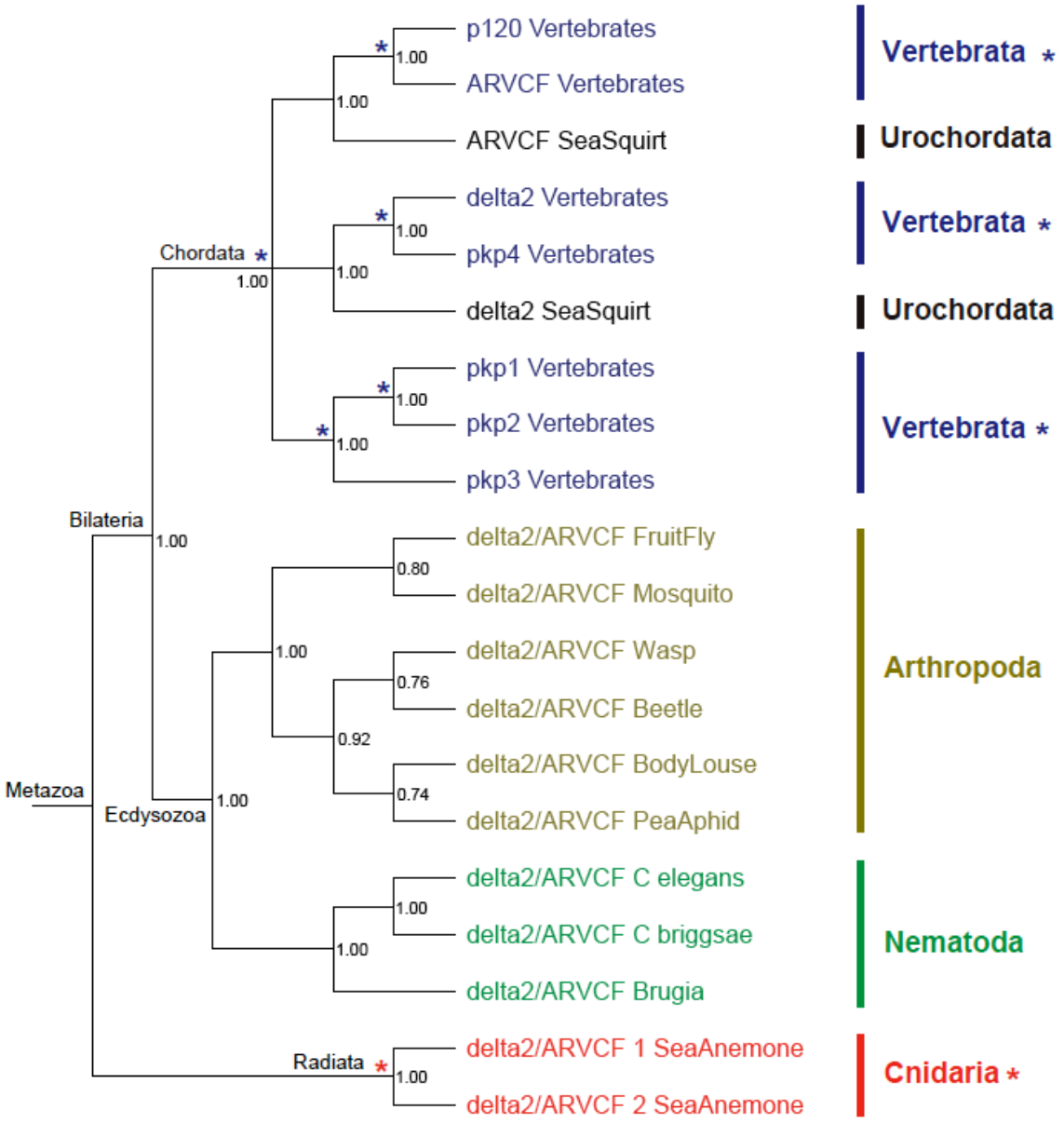


Figure 4.3: The evolution of the p120 subfamily.
The Bayesian phylogeny is supported with posterior probabilities, having sea anemone as the outgroup. Asterisks indicate duplication events. Coloring scheme is identical to Figure 4.1.

Our analyses suggest that the seven members of the p120 subfamily in vertebrates emerged by non-contemporaneous duplication events but the majority of these subfamily members (five out of seven) occurred with the origin of vertebrates. Delta2 and ARVCF diverged with the origin of chordates while p120 duplicated from ARVCF with the origin

of vertebrates and pkp4 duplicated from delta2 catenin with the origin of vertebrates. It is unclear exactly which member of the p120 subfamily pkp1, pkp2 and pkp3 duplicated from since this monophyletic group does not confidently clade with either ARVCF or delta2 from the urochordate. We can assume, however, that pkp1, pkp2 and pkp3 ultimately evolved from delta2/ARVCF because the ancestor of chordates appears to have had only one p120 member (delta2/ARVCF).

In addition to vertebrate-specific subphylum and chordate-specific phylum duplications, the p120 subfamily may have undergone species-specific duplications in Cnidaria. We cannot rule out the possibility that this duplication occurred at a higher level since only a single cnidarian species was incorporated in our analyses. Support of the former scenario, however, comes from an analysis of two truncated p120 subfamily homologs from *Hydra* (with NCBI protein identifiers 221130487 and 221131941 separately). A phylogenetic analysis groups the two sea anemone delta2/ARVCF as a monophyletic clade to the exclusion of the two delta2/ARVCF from *Hydra* that also form a monophyletic clade (not shown since the sequences are not complete). It is possible that delta2/ARVCF has undergone species-specific duplications in sea anemone and *Hydra* separately, but a more confident conclusion will depend on complete sequences from *Hydra* and/or the availability of sequences from other cnidarians.

In total, all p120 subfamily members share a common ancestor with delta2/ARVCF from non-chordates. This subfamily experienced multiple vertebrate-specific subphylum duplications, a single chordate-specific phylum duplication and possible species-specific duplications in the Cnidaria phylum.

Origins and duplications of beta subfamily members

Figure 4.4 shows the Bayesian phylogeny for beta subfamily members with beta catenin from sea anemone serving as the outgroup. The complete tree with branch lengths and NCBI sequence identifiers is shown in Supplementary figure C.1-B. Figure 4.4 shows that gamma and beta in vertebrates form a monophyletic clade (PP = 0.99) with all non-vertebrate betas positioned outside the clade. This result suggests that gamma and beta catenin in vertebrates share a common ancestor with beta in non-vertebrates, and gamma duplicated from beta with the origin of vertebrates.

In addition to a vertebrate-specific subphylum duplication, the beta subfamily has undergone species-specific and phylum-specific duplications. Figure 4.4 shows that in the Nematoda phylum, three monophyletic clades are formed by bar1, wrm1, and hmp2 separately (all having PP = 1.0), suggesting two Nematoda-specific phylum duplications. However, only two of the four species (*C. elegans* and *C. Briggsae*) studied here contain the complete three beta catenin paralogs, while *Pristionchus* and *Brugia* seem to have lost wrm1. Additionally, *Brugia* contains two beta paralogs (labelled as beta2 and beta3 arbitrarily by us) within the bar1 clade, indicating a *Brugia* species-specific duplication within the Nematoda phylum. Thus the beta subfamily has undergone a species-specific duplication in *Brugia* and two phylum-specific duplications.

For the Arthropoda phylum, two monophyletic clades (both having PP = 1.0) are formed from two individual species-specific duplications in the pea aphid and the body louse. Each group is the result of a single duplication resulting in two paralogs. Beetle also contains what appears to be a species-specific duplication but the paralogs do not group together so we cannot rule out an ancient origin for this duplication. As for the Cnidaria phylum, we have found two beta catenin homologs in sea anemone and only one

in *Hydra*. We only show one beta catenin (442 amino acids, aa) for sea anemone in Figure 4.4 since the other potential paralog (with the NCBI protein identifier 156615300) has been truncated to 298 aa, and this short length makes it hard to resolve the phylogenetic position for this paralog. The potential presence of two beta catenin homologs in sea anemone indicates a possible species-specific duplication in the Cnidaria phylum.

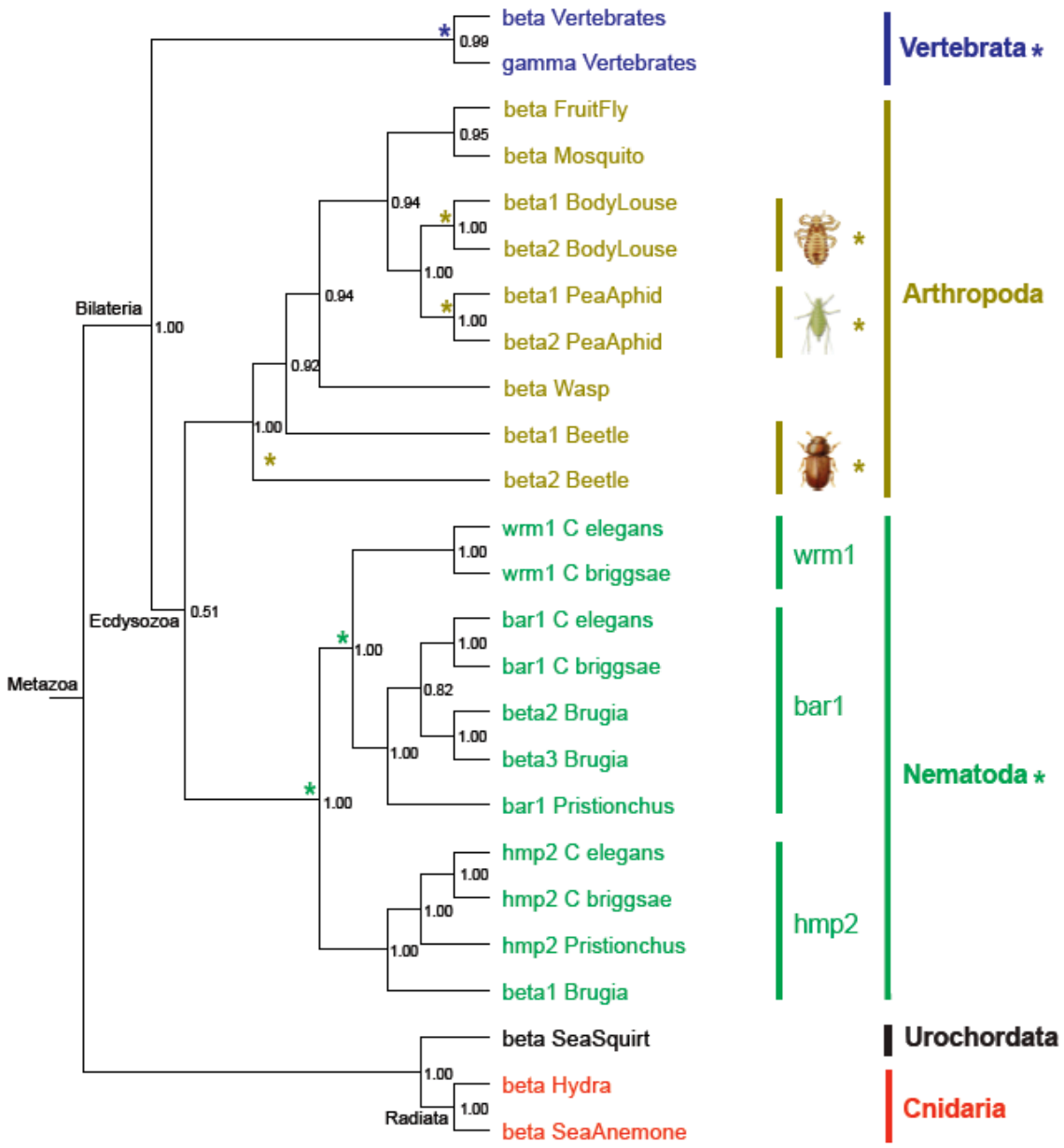


Figure 4.4: The evolution of the beta catenin subfamily.
The Bayesian phylogeny is supported with posterior probabilities, having sea anemone as the outgroup. Asterisks indicate duplication events. Coloring scheme is identical to Figure 4.1.

In total, all beta subfamily members share a common ancestor with beta catenin from non-vertebrates. This subfamily experienced a vertebrate-specific subphylum duplication, two nematode-specific phylum duplications, a species-specific duplication in the

Nematoda phylum, multiple species-specific duplications in the Arthropoda phylum and a possible species-specific duplication in the Cnidaria phylum. The conflict between the gene tree of the beta subfamily and the species tree of metazoan evolution for the position of the urochordate sea squirt is caused by a combination of slow evolution outside the nematode clade and rapid evolution inside the nematode clade (Supplementary figure C.1-B). This difference in evolutionary rates is supported by Wilcoxon rank and Kolmogorov-Smirnov tests, both with P values < 0.0001.

Origins and duplications of alpha subfamily members

The alpha subfamily has not experienced extensive duplication events similar to the p120 and beta subfamilies. Thus, the evolutionary history of this gene subfamily is easier to resolve. Supplementary figure C.1-C shows that alpha1 and alpha2 in vertebrates form two monophyletic clades (both having PP = 1.0) and together form a clade (PP = 1.0) to the exclusion of alpha3 - forming its own monophyletic clade (PP = 1.0). The two vertebrate-specific and one amniote-specific alphas form a monophyletic clade with all non-vertebrate alphas positioned outside the clade. Our analyses suggest alpha1 duplicated from alpha2 with the origin of vertebrates. It is unclear exactly which member of the alpha subfamily alpha3 duplicated from since this monophyletic group does not confidently clade with either alpha1 or alpha2 from vertebrates. We can assume, however, that alpha3 ultimately evolved from alpha2 because the ancestor of vertebrates appears to have had only one alpha member (alpha2). Our results show that the alpha subfamily experienced a vertebrate-specific subphylum duplication and an amniote-specific duplication. In total, all alpha subfamily members share a common ancestor with alpha2 from non-vertebrates.

Annotation issues for the catenin family

Our evolutionary studies have resolved annotation issues in all three catenin subfamilies (Supplementary table C.1) and have allowed us to identify nomenclature of catenin genes present before the diversification of vertebrates. Some issues are similar between subfamilies while other issues are unique to specific subfamilies.

The urochordate sea squirt contains a p120 subfamily homolog that has been annotated as 'p120'. Our analysis shows that the sea squirt sequence annotated as p120 is slightly more similar in sequence identity to ARVCF members in vertebrates (36-41%) than p120 members in vertebrates (36-40%). We therefore had to perform additional analyses to resolve this ambiguity. First, a BLASTP search using this sea squirt 'p120/ARVCF' as the query against only p120s and ARVCFs from vertebrates was carried out. This analysis demonstrated that all ARVCFs have higher BLASTP scores (bits) and lower E-values to the sea squirt sequence than the p120s from vertebrates. Second, we performed MEGA pairwise distance comparisons of the sea squirt 'p120/ARVCF' to vertebrate p120s and ARVCFs using both Dayhoff and JTT models (no. of sites = 728aa). Both models gave the same results; the distances between the sea squirt 'p120/ARVCF' and the vertebrate ARVCFs are shorter than the distances between the sea squirt 'p120/ARVCF' and the vertebrate p120s, and this difference in distance is statistically significant (P value equals 0.05) if both human ARVCF and p120 are removed from the analysis. We extended these analyses to include measures of functional divergence. We used the DIVERGE software package to identify type-II functional sites (a site that is conserved-but-different between clades) to help resolve the annotation of the sea squirt 'p120/ARVCF'. Of the 70 type-II sites identified by DIVERGE, 36% of sites (25) support the annotation of sea squirt as ARVCF while only 20% (14) support the

annotation as p120. The remaining sites do not have resolution in favor of one annotation over the other. Thus, the BLASTP, MEGA and DIVERGE analyses all suggest that the sea squirt 'p120/ARVCF' is more similiar to the ARVCF group than the p120 group.

The other p120 subfamily homolog in the urochordate sea squirt has been annotated as 'delta2'. Given the close relationship between delta2 and pkp4 (Figure 4.3 and Supplementary figure C.1-A), we applied the same analyses used for the sea squirt 'p120/ARVCF' to resolve the annotation of the sea squirt 'delta2/pkp4'. Both BLASTP and MEGA analyses suggest that the sea squirt 'delta2/pkp4' is closer to delta2 than pkp4. The MEGA result is statistically significant as supported by Wilcoxon rank and Kolmogorov-Smirnov tests, both with P values equal to 0.01. DIVERGE analysis further supported the annotation of the sea squirt sequence as delta2. Of the 88 type-II sites identified by DIVERGE, 32% of sites (28) support the annotation of sea squirt as delta2 while only 17% (15) support the annotation as pkp4. The remaining sites do not have resolution in favor of one annotation over the other. This implies that delta2 members evolved prior to the evolution of pkp4 members. Accordingly, we retain the original annotation of this sea squirt gene as delta2.

In regards to the p120 subfamily, only two p120 subfamily members are present in the urochordate sea squirt. This indicates they are the two oldest members of this subfamily. Bayesian phylogenetic analysis confirmed that all the p120 subfamily homologs in non-chordates share a common ancestor with all seven members of the p120 subfamily in vertebrates (Supplementary figure C.1-A). This phylogeny, however, does not provide conclusive evidence for which of the two oldest p120 subfamily members (ARVCF or delta2) gave rise to the vertebrate expansion of the subfamily. Therefore we

performed extensive BLASTP analyses and series of MEGA pairwise distance comparisons to infer whether delta2 or ARVCF may be the more ancient member of the p120 subfamily. Both BLASTP and MEGA analyses suggest that the p120 subfamily homologs in non-chordates are closer to delta2 than ARVCF. This conclusion holds for all vertebrates except when the urochordate sea squirt is included in the analysis, and this is probably caused by rapid evolution as seen from the long branches leading to sea squirt delta2 and sea squirt ARVCF (Supplementary figure C.1-A). The MEGA analysis, however, is not statistically significant. Further, a DIVERGE analysis did not reveal sites that could help resolve the annotation of non-chordate p120 subfamily members. As such, it is not possible to apply one particular annotation to the non-chordates over another annotation.

Our analyses point to a delta2/ARVCF origin for the entire p120 subfamily since only delta2 or ARVCF is inferred to exist in non-chordates and the diversification of this subfamily occurred with the evolution of chordates. Unfortunately, gene annotation does not follow this pattern. Numerous non-chordate homologs in the p120 subfamily have been annotated as p120, armadillo repeat protein, plakophilins or fibronectins. To resolve these discrepancies we performed three BLASTP analyses using the 'mis-annotated' homologs as queries; #1) a BLASTP search of the non-redundant protein database, #2) a BLASTP search of the human protein database, and #3) a BLASTP search of the individual genomes from which the mis-annotated homolog resides. For both analyses #1 and #2, all queries resulted in trusted top-hits annotated as p120 subfamily members. For analysis #3, only a single homolog was identified in the individual genomes of arthropods and nematodes while searches of the cnidarian genomes retrieved two homologs (as

discussed above, this duplication appears to be species-specific). We, therefore, annotated all these p120 subfamily homologs in non-chordates as delta2/ARVCF (Supplementary table C.1).

For the beta subfamily, our presence/absence and phylogenetic analyses suggest that gamma catenin is the result of a duplication event from beta catenins that took place during the origin of vertebrates and thus there should not be any gamma catenins present in non-vertebrates. One gene from *S. mansoni*, however, is annotated as gamma catenin. We performed the same series of three BLAST searches described above. These analyses suggest that this homolog is neither more similar in sequence to gamma nor beta as it appears to have experienced an episode of rapid evolution along the branch leading to *S. mansoni* (data not shown). However, since our presence/absence and phylogenetic analyses suggest that only beta is present in non-vertebrates, we infer that this gene evolved from a beta catenin and has since lost most of the sequence signatures that separate beta from gamma (it has retained enough signatures that allow us to confidently identify this as a member of the beta subfamily). Thus our analyses suggest that this gene arose from a duplication event of beta catenin in *S. Mansoni* and we re-annotated this gamma as beta catenin (Supplementary table C.1). *Aplysia californica* is the other species from the same phylum as *S. mansoni* that has whole genome (shotgun) sequence available. A BLAST search of this genome reveals only a single gene belonging to the beta subfamily. This suggests that the paralogs in *S. mansoni* arose via a species-specific duplication event. In total, our analyses suggest that beta is the most ancient member of this subfamily with gamma arising during the origin of vertebrates while other duplication events have taken place in individual species such as *S. mansoni*.

For the alpha subfamily, our presence/absence and phylogenetic analyses suggest a vertebrate origin for alpha1 catenin. Both BLASTP and MEGA analyses suggest the alpha subfamily member present in non-chordates is more similar to the alpha2 than the alpha1 from vertebrates. The MEGA results are statistically significant as determined using both Wilcoxon rank and Kolmogorov-Smirnov tests (P equals 0.01 and 0.02, respectively). The distance to the arthropod fruit fly, however, was not significant, presumably due to the short branch leading to the fly (Supplementary figure C.1-C). Our series of three BLAST analyses identified only a single alpha subfamily homolog in each of the non-vertebrates studied here. Thus all alpha subfamily homologs in non-vertebrates should be alpha2. As such, the annotated alpha1 in *Pediculus humanus corporis* (body louse) would be incorrectly annotated. We propose that members of the alpha subfamily in non-vertebrates be annotated as alpha2 (Supplementary table C.1).

Distant relatives of catenins in non-metazoans

The p120 subfamily members have been detected in all metazoans, but not in the premetazoan *M. brevicollis* or other non-metazoans. However, relatives of this subfamily have been detected in non-metazoans by our BLAST searches. These included armadillo repeat-containing protein (armc) 3, armc4, the importin alpha family, and vac8p.

The beta subfamily members have been detected in all metazoans, but not in the premetazoan *M. brevicollis* or other non-metazoans. However, several potential beta catenin homologs and beta catenin-like genes have been detected in non-metazoans by BLAST. These genes included aardvark, vac8p, arabidillo-1/-2, physcodillo-1/-2, and beta-catenin-like.

We did extensive analyses by BLASTP and Bayesian phylogenetics to determine the relationship of distant homologs of both the p120 and beta subfamilies. Our analyses show that arabidillo-1/-2 and physcodillo-1/-2 formed a well-supported clade (PP=1.0), and aardvark, vac8p, beta-catenin-like, armc3, armc4 and importin alpha all form individual monophyletic clades (PP=1.0, phylogeny not shown). Thus, none of these distant homologs grouped with the beta or p120 subfamilies in our phylogenetic analyses.

Alpha subfamily members have been detected in all metazoans but not in the premetazoan *M. brevicollis*. However, a relative of this subfamily (vinculin-like) has been detected in the premetazoan *M. brevicollis*. An additional relative (alpha-catulin) was detected by our BLAST analyses but present only in metazoans. Bayesian phylogenetic analyses showed that alpha catenins, vinculin and alpha catulin all formed three monophylogenetic clades (PP equals 0.99, 0.97 and 1.0, respectively).

Discussion

Origins and evolution for the catenin family

Despite similar names, alpha catenins are not homologous to p120 and beta catenins. This is evident from sequence and structural comparisons (Figure 4.2). Conversely, this figure shows that beta and p120 share a common ancestor as a result of a duplication event prior to the diversification of metazoans. This evolutionary relationship from sequence-based analyses is consistent with the differences in secondary structures between alphas versus p120s/betas. The p120 subfamily contains 9 ARM repeats [99] and the beta subfamily contains 12 of these ARM repeats [83]. The alpha subfamily, however, contains three vinculin homolog domains instead of ARM repeats, and belongs to the vinculin superfamily [84]. Therefore, catenin should not be called a family, since it is just a group of protins binding C-terminal of classical cadherins. For the p120

subfamily, we conclude that it is more appropriate to refer to the p120 subfamily as delta. We suspect that this change in nomenclature will be difficult to accept at first (as all name changes are) but our proposed change more accurately reflects the evolution of this subfamily and the new nomenclature is more functionally consistent for experimental biologists.

Delta2 or ARVCF is the oldest member of the p120 subfamily according to our analyses but we did not detect it in the premetazoan *M. brevicollis*. This suggests a metazoan origin for the p120 subfamily. Unlike the p120 subfamily, we have detected the relatives of this subfamily beyond metazoans, such as armc3, armc4, the importin alpha family, and vac8p. This suggests that these armadillo repeat proteins share a common ancestor. This can be further inferred by their sequence and structural conservations though they have diverse functions in different kingdoms including animals, fungi and plants [33].

Beta catenin is the oldest member of the beta subfamily according to our analyses but we did not detect it in the premetazoan *M. brevicollis*. This indicates a metazoan origin for the beta subfamily. Previous studies also suggested the potential metazoan origin of beta catenin and its involvement in cell adhesion for multicellular organisms [100]. Several types of genes in non-metazoans, however, have been considered potential beta catenin homologs [33] and thus some have been annotated as beta catenin or beta catenin-like. These genes included aardvark in amoeba, vac8p in fungi and plants, arabidillo-1/-2 and physcodillo-1/-2 in plants, and beta-catenin-like in both animals and plants. The aardvark gene in amoeba participates in adherens junctions and cell signaling, and it may be the amoeba's version of a precursor to beta catenin [101]. Vac8p has 11 ARM repeats

and this is close to the 12 ARMs that compose beta catenin [102]. None of these potential beta homologs grouped with the beta subfamily in our phylogenetic analyses, thus supporting our hypothesis of a metazoan origin for the beta subfamily.

Our analyses also revealed a significantly higher mutation rate within the Nematode clade for beta catenin paralogs specific to this clade. The high rate can be correlated to experimental work showing subfunctionalization of celluar adhension and transcriptional activation (signalling) among the three paralogs in *C. elegans* revealed by their protein-binding partners [103]. It remains to be determined whether this is due to relaxation of selective constraints or due to positive selection.

Alpha2 catenin is the oldest member of the alpha subfamily according to our analyses but we did not detect it in the premetazoan *M. brevicollis*. This suggests a metazoan origin for the alpha subfamily. We did however detect a vinculin-like gene in the premetazoan *M. brevicollis*. This suggests that the alpha subfamily duplicated from vinculin-like genes after the separation of metazoans/premetazoans but prior to the diversification of metazoans.

In summary, our analyses did not identify any p120, beta or alpha subfamily members in the premetazoan *M. brevicollis*. We confirmed this absence by extending our analyses to other non-metazoans such as the amoeba *D. discoideum*, the protist *C. parvum*, three fungi (*Aspergillus*, *Saccharomyces*, and *Schizosaccharomycetes*), three unicellular algae (*Ostreococcus*, *Chlamydomonas*, and *Thalassiosira*), and the plant *Arabidopsis*. Again, none of these non-metazoans contained a member of the catenin family. We conclude that the catenin family arose during the origin of metazoans while some other armadillo

repeat proteins and vinculin-like proteins were present prior to the evolution of metazoans.

Evolution-related physiology of the catenin family

The members of the three catenin subfamilies emerged at different points during metazoan evolution. It would be interesting to correlate gene evolution of catenins to the cellular physiology associated with the functional divergence of this gene family.

Our results suggest that either delta2 or ARVCF is the oldest member of the p120 subfamily - one evolved with the origins of metazoan while the other evolved with the origins of chordates. Pkp4, p120, pkp1, pkp2, and pkp3 show vertebrate origins and this is consistent with the conclusions of other analyses [89]. Functional divergence between delta2 and ARVCF in chordates is highlighted by delta2's neuron-specific expression while ARVCF is ubiquitously expressed [82]. Both ARVCF and p120 are ubiquitously expressed in epithelial tissue [82], but their expression is mutually exclusive [104] and have complementary distributions in epithelial adherens junctions as interpreted from BioGPS [96]. Our evolutionary studies show that p120 duplicated from ARVCF during the origin of vertebrates. The origin of p120 members appears to coincide with the vertebrate origin of p120's binding partner Kaiso [89], a nuclear factor participating in signaling pathways [105]. This interaction might relate to epigenetic transcriptional regulation and Wnt signaling modulation via methyl-CpG islands [43] and the transcription factor TCF [6], participating in vertebrate-specific development and transcriptional regulation [89]. The nuclear signaling functions of ARVCF have not yet been identified but it is known that it interacts with the novel protein Kazrin localized in both the cytoplasm and nucleus [81]. P120 is unique among the members of the p120

subfamily because it interacts in the nucleus with the nuclear transcription repressor Glis2 [106]. This demonstrates that p120 has diverged from ARVCF due to its unique interaction with Glis2 and its vertebrate-specific interaction with Kaiso, but p120 still shares functional redundancy with ARVCF since they can rescue one another's null mutants [107]. Another unique feature of p120 is the absence of the C-terminal PDZ binding domain. The lack of this domain may enable p120 and its binding cadherins to evolve with more flexibility in a PDZ-independent manner [89]. In support of this view is the observation that the function of delta2 catenin in spine density regulation is independent of cadherins but dependent on an interaction with PDZ-domain containing proteins while p120 regulates spine density using an alternative mechanism that depends on Rho GTPases [108].

Both p120 and ARVCF, but neither delta2 nor pkp4, have different isoforms that result from alternative splicing at the 5' portion of the transcript [104, 109]. Translating the conserved N-terminal coiled coil domain during such isoform switching, and in coordination with the dynamic cadherin switching (e.g., E-cadherin to N-cadherin), occurs during the transformation from epithelial and other sessile cell types to mesenchymal (e.g., fibroblasts) and other motile cell types (e.g., neurons) [82]. This process might be important in development, wound healing and cancer [89].

Pkp4 duplicated from delta2 during the origin of vertebrates and both have neuron-specific expression patterns, however this pattern of expression is reciprocal according to BioGPS. Delta2 is predominantly expressed in the brain while pkp4 is expressed in the brain but also ubiquitously expressed in other tissues. Pkp4 differs from delta2 in that pkp4 participates in desmosome formation in addition to adherens junctions according to

numerous experimental reports [38], however one study failed to verify this dual functionality [110]. Pkp1, pkp2, and pkp3 are vertebrate-specific and are well known for their functions in desmosomes instead of adherens junctions [111]. The origins of these plakophilins appear to coincide with our assumption of the origin of desmosomes in vertebrates since our analyses reveal that other desmosome-specific genes are found only in vertebrates (further discussion is provided below).

Our results suggest that beta is the oldest member of the beta subfamily since it is present in all metazoans while gamma is a more recent acquisition since it is found only in vertebrates. Both beta and gamma are ubiquitously expressed at basal levels but gamma is uniquely highly-expressed in some tissues such as tongue. Beta and gamma both function in the formation of adherens junctions however gamma has functionally diverged from beta since gamma can also participate in desmosome formation [85].

Gamma (also called plakoglobin) and the four plakophilins (discussed above) are all components of the desmosome [85] and all have vertebrate origins according to our analyses. This suggests to us that the desmosome evolved in conjunction with the origin of vertebrates. To further validate this hypothesis we performed computational analyses of other non-catenin gene families involved in desmosome formation. These included both desmosomal cadherins desmogleins and desmocollins, as well as desmoplakins and thus we confirmed that the presence of genes involved in desmosome formation is limited to vertebrates and not observed in non-vertebrates. All of our results are consistent with the observation that desmosome-containing tissues (e.g., certain cardiac and skeletal muscles [112], keratin-containing hair [113], and others) have evolved with or after the origin of vertebrates.

Our results suggest that alpha2 is the oldest member of the alpha subfamily since it is present in all metazoans while alpha1 is a more recent acquisition since it is found only in vertebrates. Alpha1 and alpha2 display both similar and reciprocal expression and distribution patterns in the dorsal root ganglia and spinal cord at the lumbar level where sciatic nerves originate [114]. The two are similar in that they both display brain-specific expression patterns however they display reciprocal expression patterns in terms of the specific types of tissues in the brain [115]. Unlike alpha1, alpha2 functions by binding the nuclear transcription repressor ZASC1 [116]. Alpha2 knockouts in mice mainly affect the nervous systems by causing cerebellar deficient folia [84]. Alpha1 knockouts in mice are lethal at the blastocyst stage but its knockout mainly affects the epithelial tissues such as skin [84].

Alpha3 is the most recently evolved member of the alpha subfamily since it is found only in amniotes. The observation that alpha3 and alpha2 share identical exon-exon boundaries [117] suggests to us that alpha3 evolved via a gene duplication of alpha2 coinciding with the origin of amniotes. Unlike alpha1 and alpha2, alpha3 interacts with pkp2 in the area composita (a hybrid adherens junction in the heart muscle), yet co-expresses with alpha1 at intercalated discs of cardiomyocytes [118]. We could not find a physiological link between the amniote-specific alpha3 and the observation that it is highly expressed in the testis and participates in adherens junctions between sertoli and germ cells, highly expressed in the testis interstitial tissue or highly expressed in peritubular myoid cells of the testis [119] because the above cells or tissues are also found in the non-amniote fish. However, alpha3 interacts with 1-adadin and this interaction, along with a truncated isoform of alpha3 present in the testis [120], may be

unique to spermatogenesis in amniotes. No alpha3 knockout experiments have been performed in mice [84].

Overall, delta2, alpha2 and N-cadherin are present in nearly all metazoans and all are expressed in neural-specific tissues. Their vertebrate-counterparts p120, alpha1 and E-cadherin, however, are mainly expressed in epithelial tissues in addition to neural tissues. Extensive co-evolution exists between the catenin family and their binding partners; these include p120 with Kaiso, delta2/alpha2 with N-cadherin, p120/alpha1 with E-cadherin, and pkps/plakoglobin (gamma catenin) with desmosomal cadherins.

Conclusion

The p120, beta and alpha subfamilies are present in all metazoans, but none of the subfamilies are present in non-metazoans. This indicates a metazoan origin for the catenin family. Each catenin subfamily has undergone duplications at different levels (species-specific, subphylum-specific or phylum-specific) and to different extents (in the case of the number of homologs).

Vertebrate-specific duplications occurred in all three subfamilies: p120, pkp4, pkp1, pkp2, and pkp3 in the p120 subfamily; gamma in the beta subfamily; and alpha1 in the alpha subfamily. All three subfamilies had extensive annotation issues and these have been effectively resolved by our studies. We anticipate that this resolution in combination with our evolutionary analyses will make it easier for experimental biologists to correlate diverse catenin physiologies to interesting evolutionary innovations such as the development of multicellularity and the role of adhesion during this development and the evolution of terrestrial organisms and their ability to prevent desiccation.

Abbreviation

P120, p120 catenin; beta, beta catenin; alpha, alpha catenin; ARVCF, Armadillo Repeat protein deleted in Velo-Cardio-Facial syndrome; pkp, plakophilin; aa, amino acids; PP, Posterior Probability; ARM, armadillo; MCMC, Markov chain Monte Carlo; PSRF, Potential Scale Reduction Factor.

Acknowledgments

This work was supported by the Georgia Institute of Technology and a NASA grant to Eric A. Gaucher.

CHAPTER 5

ANCESTRAL SEQUENCE RECONSTRUCTION OF THIOREDOXIN TO UNDERSTAND ANCIENT ENVIRONMENTS [121]

Abstract

The recent accumulation of DNA sequence data, combined with advances in evolutionary theory, computational power and DNA synthesis technology, have promoted the development of Ancestral Sequence Reconstruction (ASR) methods. By ASR, ancient sequences could be reconstructed computationally by virtue of a comparison among sequences of related genes found in modern organisms and then subsequently experimentally resurrecting these ancient proteins in the laboratory to measure their properties and functions. We performed ASR to a ubiquitous enzyme present in all three domains of life called thioredoxin, and reconstructed ancient genes including the last common ancestors of all bacteria, all eukaryotes, all archaea, and all eukaryotes and archaea. This reconstruction allows us to trace back as far as ~ 4 to ~1.4 billion years (Gyr), and serves as a fundamental step for following experimental investigations in biochemical functions of ancient thioredoxins to infer paleoenvironments.

Introduction

Experimental palaeogenetics and palaeobiochemistry provide an opportunity to investigate in the laboratory the molecular history of modern organisms. The study of

resurrected proteins can also reveal valuable information regarding the environmental adaptation of extinct forms of life which may be associated to climatic, ecological and physiological alterations [34-37]. Despite numerous experimental examples of protein resurrection in the laboratory [122], the majority of the resurrected proteins have provided a journey back in time of no more than 200-300 millions years (Myr) [33, 39, 122]. Consequently, many hypotheses about ancient life remain untested, especially in time periods in which dramatic changes in biological systems occurred [123]. This time traveling is largely limited by the ambiguity in the historical models used for ancestral sequence inference. For instance, uncertainties in databases, sequence alignments, failures in evolutionary theories and uncertainty in the construction of phylogenetic trees are common sources of ambiguity [122]. Nonetheless, several approaches are commonly used to overcome these limitations and some examples of resurrected proteins from the Precambrian supereon (4.5-0.5 Gyr ago) have been reported [122, 124]. A prominent case is the resurrection of the Elongation Factor (EF) from ancient bacteria 0.5 to 3.5 Gyr old [39, 125]. The ancestral EFs demonstrated that ancient bacteria lived in a hot environment and also revealed a clear correlation between the thermostability of the proteins and the temperature of ancient oceans as inferred from geological records.

These pioneer studies paved the way to formulate countless questions about ancient organisms and biomolecules close to the time of the origin of life. For instance, little is known about how the chemistry of primitive enzymes arose and how the environmental conditions affected the evolution of their chemistry. Certainly, these enigmas cannot be resolved by examining fossil records. In an effort to understand the evolution of enzymatic reactions we have reconstructed thioredoxin enzymes (Trx) from extinct

organisms that lived in the Precambrian. Thioredoxins belong to a broad family of oxidoreductase enzymes that are ubiquitous in all living organisms [126], thus our ancestral sequence reconstruction of thioredoxin covers all three domains of life including bacteria, archaea, and eukaryotes, tracing back as far as ~ 4 to ~1.4 Gyr. A series of computationally reconstructed Trx were resurrected and tested in the laboratory for their enzymatic activities and temperature/PH preferences, to infer environments on the early Earth.

Methods

Phylogenetic analyses

We retrieved 203 thioredoxin sequences covering all three domains of life from GenBank with GI numbers recorded (Supplementary Note D.1). Sequences were aligned using MUSCLE [63] and further corrected manually. The phylogenetic analysis was carried out by the minimum evolution distance criterion with 1,000 bootstrap replicates using PAUP* 4.0 beta [127].

Ancestral sequence reconstruction

Ancestral sequences were reconstructed using PAML 3.14 [98] and incorporated the gamma distribution for variable replacement rates across sites. For each site of the inferred sequences, posterior probabilities were calculated for all 20 amino acids. The amino acid residue with the highest posterior probability was then assigned at each site.

Results and discussion

Phylogenetic tree of thioredoxin

Phylogenetic tree of 203 Trx sequences from three domains of life was constructed (Figure 5.1). Seven internal nodes of interest are highlighted with red arrows, including last bacterial common ancestor (LCBA), last archaeal common ancestor (LACA),

archaea/eukaryota common ancestor (AECA), last common ancestor of cyanobacterial and deinococcus/thermus groups (LPBCA, origin of phosynthetic bacteria), last eukaryotic common ancestor (LECA), last common ancestor of γ-proteobacteria (LGPCA) and last common ancestor of animals and fungi (LAFCA). Owing to the extensive number of extant Trxs sequences available, we also have constructed a highly articulated phylogenetic tree representing the three domains of life and the interested ancestral nodes with divergence time labeled (Figure 5.2A).

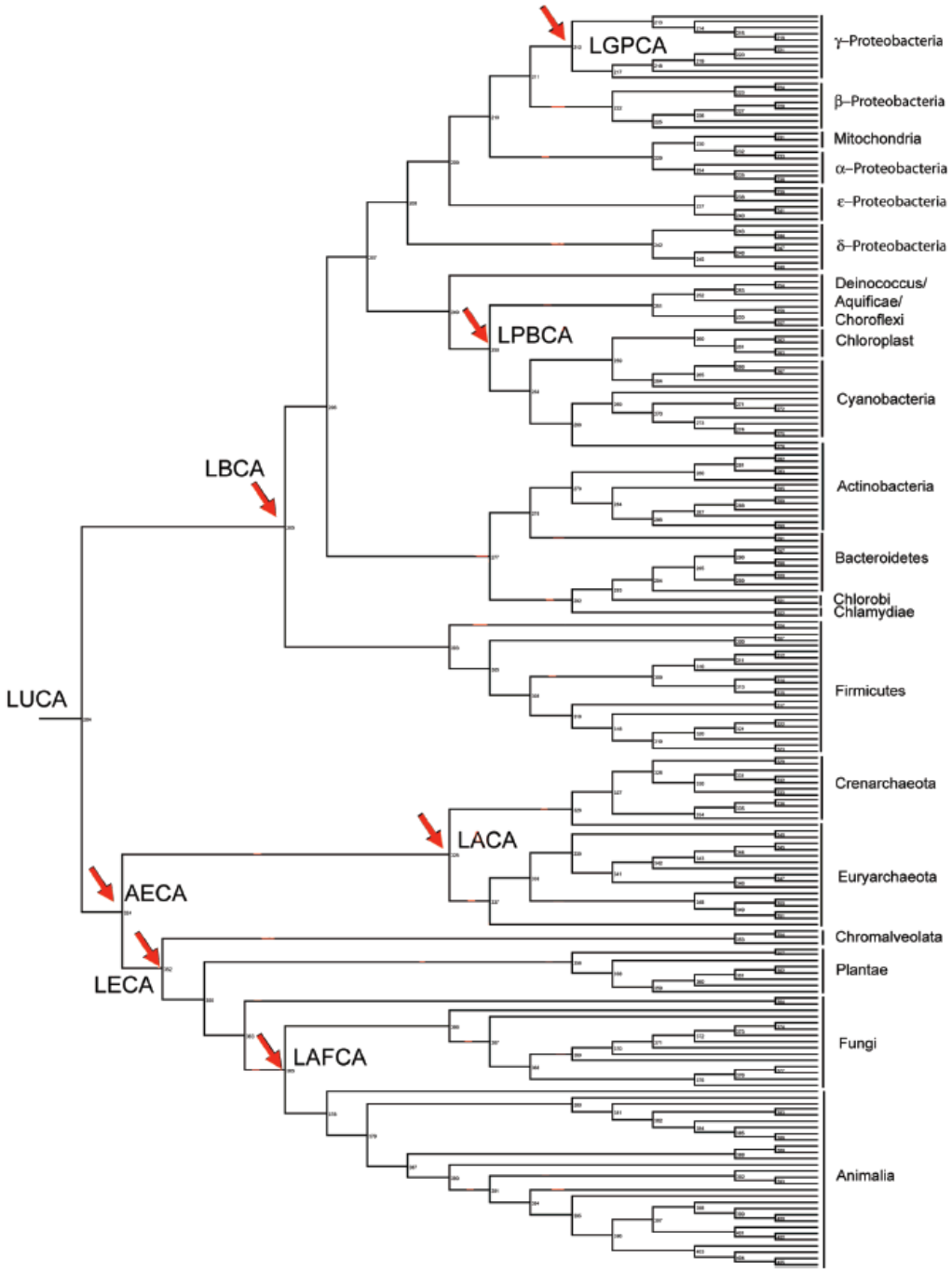
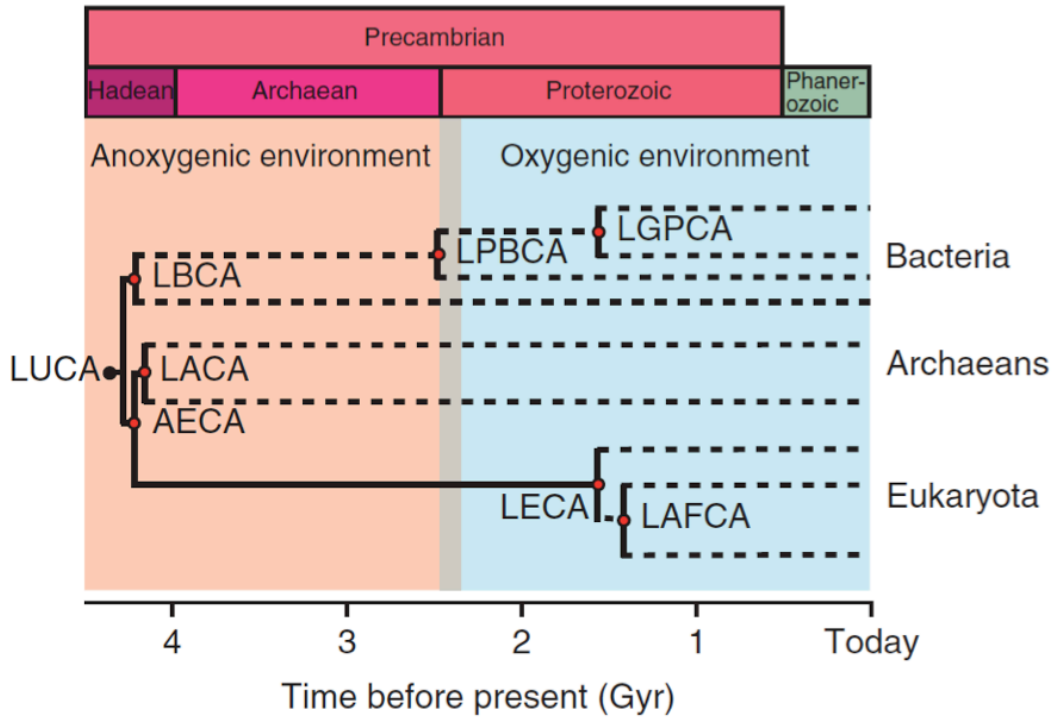


Figure 5.1: Phylogenetic Tree used for the ancestral sequence reconstruction of Trx enzymes.
A total of 203 sequences were used and the internal nodes were labeled in number. The nodes of interest are highlighted with red arrows, including last bacterial common ancestor (LCBA), last archaeal common ancestor (LACA), archaea/eukaryota common ancestor (AECA), last common ancestor of cyanobacterial and deinococcus/thermus groups (LPBCA, origin of phosynthetic bacteria); last eukaryotic common ancestor (LECA), last common ancestor of γ-proteobacteria (LGPCA) and last common ancestor of animals and fungi (LAFCA).

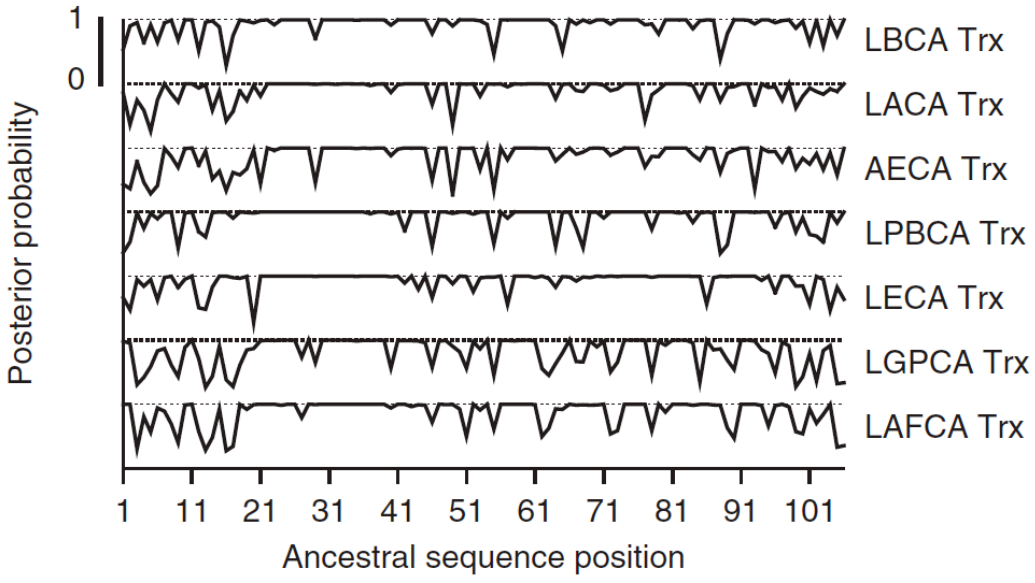
Reconstruction of ancestral Trx enzymes

From the phylogenetic tree of Trx, we sampled several biologically relevant internal nodes for sequence reconstruction (Figure 5.2A). Divergence dates estimates were applied to nodes in the tree [128]. In particular, we reconstructed seven Trx enzymes belonging to the last bacterial common ancestor (LBCA), the last archaeal common ancestor (LACA) and the archaea/eukaryota common ancestor (AECA). These organisms are thought to have inhabited Earth in Hadean and/or early Archaean 4.2-3.5 Gyr ago after diverging from the last universal common ancestor (LUCA) [36]. We also selected the node corresponding to the last eukaryotic common ancestor (LECA), that lived in the Proterozoic eon, ~1.60 Gyr ago, by the time that oxygen levels were elevated [129]. Two other internal nodes in the bacterial lineages were selected, the last common ancestor of photosynthetic bacteria (LPBCA) which existed 2.50 Gyr ago, and the last common ancestor of γ-proteobacteria, 1.61 Gyr old (LGPCA). Finally, we also chose the last common ancestor of animals and fungi (LAFCA) that lived 1.37 Gyr ago.

The sequences of seven ancient Trx enzymes were reconstructed using statistical methods based on maximum likelihood [39, 130]. For a given node in the tree, we calculated posterior probability values for all 20 amino acids considering each site of the inferred sequence. These values represent the probability that a certain residue occupied a specific position in the sequence during evolution. The posterior probabilities were calculated on the basis of an amino acid replacement matrix [131]. The most probabilistic ancestral sequence (M-PAS) at a specific node was then reconstructed by assigning to each site the residue with the highest posterior probability. Figure 5.2B shows the posterior probability distribution of the inferred amino acids across 106 sites for the selected sequences.



A:



B:

Figure 5.2: Phylogenetic analysis of Trx enzymes and ancestral sequence reconstruction.

A; Schematic phylogenetic tree showing the geological time in which different extinct organisms lived (see text). Dashed lines represent further bifurcations. Divergence times are compiled from multiple sources and are summarized. The figure indicates the global environment, although both aerobic and anaerobic organisms are found in modern environments.

B: Posterior probability distribution of the inferred amino acids across 106 sites for the interested internal nodes. The inferred amino acid at each site for the interested internal node is the residue with the highest posterior probability.

The computationally reconstructed ancestral thioredoxin genes were synthesized and expressed in E. coli cells (Epoch Biolabs). Chemical mechanisms of synthesized ancient thioredoxins were observed by our collaborators from Julio M Fernandez and Jose M Sanchez-Ruiz's groups, as well as preferences of these ancient genes to temperature and PH. The preferred temperature and PH by ancient genes indicates a hot and acidic primordial environment. Although Trx enzymes have maintained their reductase chemistry unchanged, they have adapted over 4 Gyr to the changes in temperature and ocean acidity that characterize the evolution of the global environment from ancient to modern Earth.

Conclusion

We computationally reconstructed ancestral sequences for a ubiquitous enzyme thioredoxin that covers all three domains of life. This reconstruction traces back as far as ~ 4 to ~1.4 billion years (Gyr) to the last common ancestors of all bacteria, all eukaryotes, all archaea, and all eukaryotes and archaea. Our computational reconstructed ancient enzymes were validated experimentally in biochemical functions and preferred environments. Experimental results show that ancient genes possess the enzymatic activities and prefer a hot and acidic environment. Therefore, ancestral resurrection of universal enzymes can be a powerful tool toward understanding the origin and evolution of life on Earth.

Abbreviation

ASR, Ancestral Sequence Reconstruction; Myr, millions years; Gyr, billion years; EF, Elongation Factor; LACA, last archaeal common ancestor; AECA, archaea/eukaryota

common ancestor; Trx, thioredoxin enzymes; M-PAS, most probabilistic ancestral sequence.

Acknowledgments

This work was supported by NASA Astrobiology (Georgia Institute of Technology) and NASA Exobiology to Eric A. Gaucher.

CHAPTER 6

ANCESTRAL GENOME RECONSTRUCTION AND MINIMAL GENOME OF MYCOPLASMAS

Abstract

To date, most Ancestral Sequence Reconstruction (ASR) studies have focused on the resurrection of single genes, such as EF-Tu, thioredoxin, etc. The Venter Institute's accomplishment of determining essential genes in a minimal bacterium, synthesizing a whole genome, and transplanting a synthesized genome from one species to another species has set the stage to substantially extend ASR to ancestral genome reconstruction (AGR). We applied various computational and evolutionary tools to reconstruct the gene content of an ancestral *Mycoplasma* genome. We performed pairwise proteomic comparisons to identify genome content (gene's presence/absence) of an inferred ancestor and to understand genome evolution by linking genotypes and phenotypes. We compared the reconstructed genome content of the ancestor to two hypothetical minimal genomes of *M. genitalium* and *M. pulmonis* and discuss gene content differences. We also reconstructed the synteny orders in the ancestor to understand genome evolution via genome structure and genome rearrangement mechanisms. Overall, ancestral genome reconstruction provides insights into genome evolution and also provides a gene list that can be utilized to create an ancient or engineered organism by the synthetic biology community.

Introduction

The J. Craig Venter group successfully synthesized a Mycoplasma genome as part of their Minimal Genome Project [132-134]. This is the first example of transplanting a synthetic genome into a recipient bacteria cell, creating a viable and self-replicating cell. Currently, this genomic manipulation technique has been exploited by Synthetic Genomics Inc. [http://www.syntheticgenomics.com/index.html] to create a form of algae that can use carbon dioxide to make energy-resourceful fuels (biofuels) and to efficiently develop vaccines. It has been suggested that genome synthesis technology can also be exploited to create a 'minimal cell' containing only genes essential for life when supplemented with sufficient nutrients [132-134]. Furthermore, genome synthesis and transplantation technology may make it possible to reconstruct and resurrect ancient life in the laboratory.

Ancestral sequence reconstruction has been applied to genes, such as EF-Tu [39], thioredoxin [121], and alcohol dehydrogenases [35], to help understand paleoenvironments and molecular mechanisms of adaptation. Our goal here is to extend reconstruction studies to the whole-genome level with the intention of reconstructing an ancient genome from Mycoplasma species in hopes of aligning our computational inference of an ancestral genome to the technology developed by the Venter Institute [135]. Thus, we focused our reconstruction on the last common ancestor of Mycoplasma species *M. mycoides* and *M. capricolum* used by the Venter Institute.

These organisms belong to the Mollicutes class in the bacterial domain and have distinct properties such as lack of a cell wall, small genome size, low GC content, sterol requirement, UGA-encoding tryptophan (Trp), host specificity, pathogenicity and rifampicin resistance [130-132]. A phylogeny of Mollicutes is shown in Figure 6.1, with

biological traits mapped according to Mollicute evolution [136-143]. Mollicutes evolved from their walled relatives Clostridium, Bacillus and Streptococcus through the loss of both the cell wall and an extensive number of genes [130, 137, 140]. Mollicutes then diverged into Acholeplasma-Anaeroplasma-Phytoplasma (AAP) [139] and Spiroplasma-Entomoplasma-Mycoplasma (SEM) groups [141] with UGA-encoding Trp emerging within SEM group [130, 131, 137, 141]. The SEM group further diverged into Spiroplasmataceae-Entomoplasmataceae (SE) and Mycoplasmatales (M) groups [140]. Mollicutes are often the focus of minimal genome studies due to the small size of their genomes.

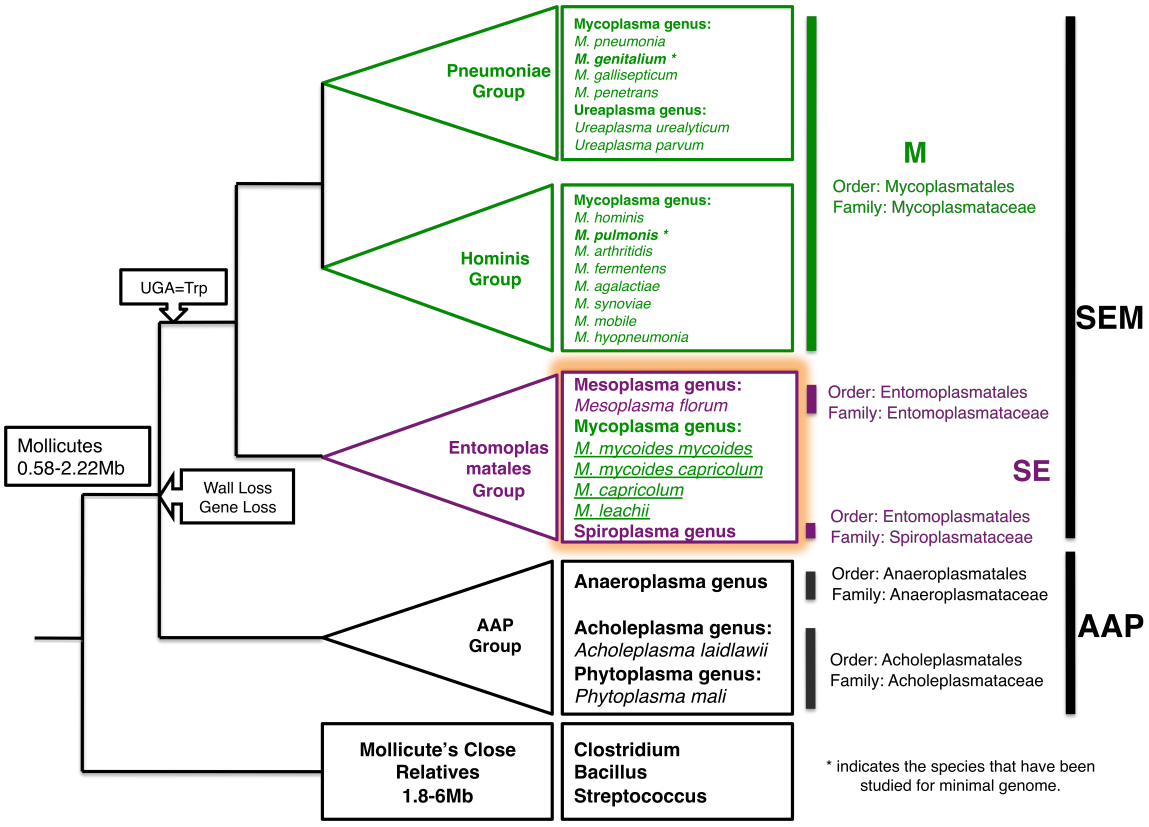


Figure 6.1: The evolution and taxonomy classification of Mollicutes.
UGA is the stop codon; Trp: Tryptophan; AAP: Acholeplasma-Anaeroplasma-Phytoplasma [139]; SEM: Spiroplasma-Entomoplasma-Mycoplasma [141]. The ranges of genome size (megabase - Mb) were provided for species in 'Mollicutes' and 'Mollicute's Close Relatives'. Representative species with the whole genome sequence available in public databases were included for Mollicutes. The green color highlights species belong to Mycoplasmatales order. The species in the mycoides cluster that we used for ancestral genome reconstruction were underlined. The purple color highlights the SE group, which forms a monophyletic clade together with the mycoides cluster evolutionarily, in a contradiction to taxonomy.

The concept of minimal genomes is over a decade old and its goal is to make a viable and self-sustaining cell encoded by a minimal genome [144]. The first steps towards this goal consisted of computational analyses that attempted to identify the set of core genes common between bacteria *Mycoplasma genitalium* and *Haemophilus influenzae.* This study suggested that the core set of genes contained 256 genes with the majority involved in housekeeping functions [145]. The next steps towards this goal consisted of experimental studies that exploited gene disruption techniques to individually turn off

genes in a species. These studies are informative in the sense that they provide statements about the effects that particular genes have on an organism's survival. However, these techniques are limited in that they cannot target all genes in a genome. Our goal is to utilize evolution-based techniques to provide insight into the evolution of Mycoplasma genomes by inferring an ancient genome and also to compare/contrast our ancient genome to minimal genomes previously reported.

Two Mycoplasma species (*M. genitalium* and *M. pulmonis*) have been previously exploited experimentally using transposon insertions to infer essential genes necessary to support a minimal genome [146, 147]. 381 and 310 genes were identified, respectively, and assumed to be essential genes. We refer to these hypothetical genomes as Mg381 and Mp310 for short. One potential limitation, however, associated with current experimental techniques is that transposons are highly unstable and are known to transpose between parental and offspring cells during the experiments with high frequency [147]. This lack of robustness in the system makes it difficult to definitively know if a gene is non-mutable or not. Further, these experiments are only carried out for one generation of growth so genes necessary for long-term viability may be mis-identified as expendable.

In addition to experimental limitations, the presence of functional redundancy within a genome can also restrict the ability to infer a minimal genome. Functional redundancy can result from either paralogous genes or nonorthologous gene displacements. For instance, if two paralogs perform identical or redundant function, then these aforementioned techniques may fail to identify essential genes because the transposon repression of one gene is compensated by the functionality of its paralog, and vice versa. As such, neither paralog would be present in a minimal set of genes. This limitation is

discussed in previous work but a suitable solution has not yet been implemented to manage functional redundancy [146-148].

These limitations highlight the potential shortcomings of current approaches towards the inference of a minimal genome. We anticipate that previous studies have provided a solid foundation of genes on which to build but that these studies have underestimated the true numbers of genes required for viability of a Mycoplasma organism.

Our approach to manage the underestimation of essential genes is to exploit the concept of ancestral genome reconstruction to computationally infer the genome content of an extinct Mycoplasma. We anticipate that inferring an ancient Mycoplasma genome will generate a more reliable list of essential genes because 1) the ancient Mycoplasma presumably existed at some point in evolutionary time, whereas the minimal genome is a hypothetical notion that may be biological inconsistent; 2) the ancient genome can manage functional redundancy contained within paralogous genes (as seen with our previous work with alcohol dehydrogenases) or nonorthologous genes; and 3) the ancient genome contains not only core genes but also non-core genes that are also essential in creating a viable self-replicating cell.

The computational reconstruction of any ancestral genome must consider both small- and large-scale genome structures. Small-scale structure includes sequence substitutions and short sequence insertions and deletions (indels). Large-scale structure includes genome content such as gene gain/loss (due to horizontal gene transfer – HGT, or gene duplications) and genome rearrangements such as translocations, transpositions, and inversions [149, 150] (Figure 6.2A). Large-scale genome reconstruction consists of inferring the content (such as gene presence/absence) and arrangement (gene

order/synteny) of an ancestral genome [149, 151, 152] (Figure 6.2B). Our current work is focused on reconstructing large-scale genome content and synteny while future work will focus on small-scale reconstruction.

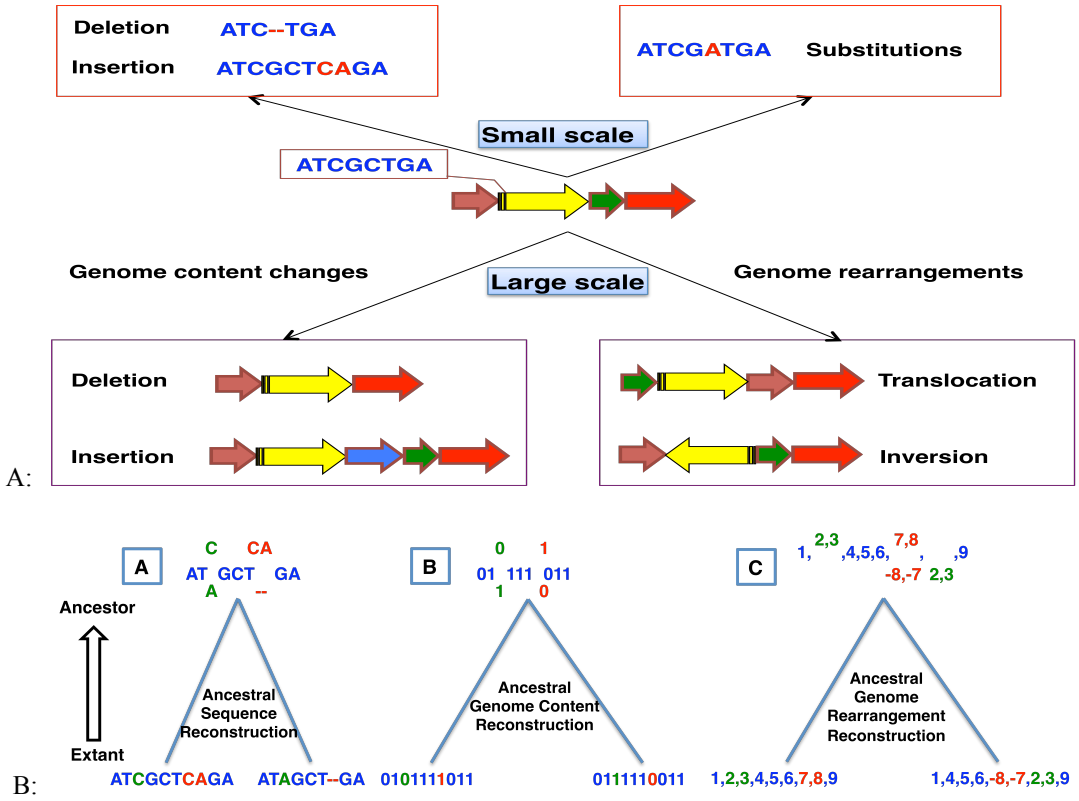


Figure 6.2: Genome evolution and ancestral genome reconstruction in small- and large-scales.
A: Genome evolution. Small-scale evolution includes sequence substitutions and short sequence indels. Large-scale evolution includes genome content changes (insertions and deletions) and rearrangements (translocations and inversions).

B: Ancestral Genome Reconstruction in small- and large-scales. B-A, Reconstruction of small-scale sequence (DNA or amino acid) substitutions and indels. The red color indicates indels, and the green color indicates point mutations. B-B, Reconstruction of large-scale genome content. The series of numbers represents a list of genes/syntenies. Gene/synteny presence and absence are characterized by 1 and 0, respectively. B-C, Reconstruction of large-scale genome rearrangement. The ordered numbers represent conserved genes/syntenies along a particular genome. The red color indicates inversions, and the green color indicates translocations.

We present our work as a first step towards the synthesis of an ancient genome. We have initially focused on computationally inferring the gene content of an ancient Mycoplasma. Future studies will focus on the gene order and intergenic regions of the ancient organism and then possibly synthesis of the ancient genome at some point in the future. If an ancient resurrected bacterial genome is synthesized and transplanted into a modern Mycoplasma cell successfully, it may be possible to synthesize an ancient organism as opposed to merely reanimating one [149]. Such a genome may also have potential applications in synthetic biology, particularly biomedicine and bioenergy.

Methods

Data sources and phylogenetic analyses

Our analyses contained four Mycoplasma species from the mycoides cluster having whole-genome sequences available in the NCBI database; two *M. mycoides* subspecies *M. mycoides mycoides PG1* (MSC) and *M. mycoides capricolum* (MLC), *M. capricolum* (MCAP), and *M. leachii* (MSB) [153]. The NCBI accession numbers for MSC, MLC, MCAP, and MSB are NC_005364.2, NC_015431.1, NC_007633.1, and NC_014751.1 respectively.

To identify an outgroup for the mycoides cluster, whole-genome sequences from non-mycoides-cluster Mollicutes were downloaded from public databases and combined with mycoides genomes to perform four evolutionary-based analyses. The first evolutionary-based analysis involved constructing a phylogenetic tree based on the 16S rRNA gene sequences extracted from the entire set of Mollicute genomes. Sequences were aligned using MUSCLE program [63] and then manually modified. Phylogenetic analyses utilized the parsimony (1000 bootstrap replicates with heuristic search) and distance (minimum evolution) (10000 bootstrap replicates with neighbor joining search)

optimality criteria in PAUP [127], and both optimality criteria were implemented with heuristic search and tree-bisection-reconnection 'TBR' branch swapping. Phylogenetic analysis also utilized the Bayesian algorithm in MrBayes 3.1.2 [79] with parameters from the DNA model GTR+G selected by jModeltest [2].

The second evolutionary-based analysis determined the presence/absence of Cluster-of-Orthologous-Groups (COGs) [154] in the Mollicute species and then used this information to group species according to the binary patterns of COGs. These binary presence/absence patterns were then used as input for phylogenetic analysis in PAUP as described above.

The third evolutionary-based analysis determined the order of locally collinear blocks (LCBs) or syntenies between the genomes (MAUVE 2.3.1 [155]) and then used this information to generate a distance matrix as input for phylogenetic analysis using the GRIMM genome rearrangement model [156].

The fourth evolutionary-based analysis used the complete proteomes of the Mollicute species to conduct pair-wise BLASTP analyses [157] and then used this information to generate a distance matrix that served as input for phylogeny construction [158]. For each pair of species, all bidirectional/reciprocal best hits (BBH) were identified by reciprocally BLASTing each of the two proteomes with e-value cutoff as 10^{-10}. The sequence identities between all BBHs were averaged and then this average served as the distance between any two pairs of proteomes. The NEIGHBOR program in PHYLIP [159] was used to construct a phylogenies using the distance matrix as input and utilizing the neighbor joining clustering algorithm.

Ancestral genome content reconstruction:

Proteome-wide pair-wise BLASTP (e-value <0.01) was performed on the five genomes MSC, MLC, MCAP, MSB, and *Mesoplasma florum* (Mf) as the outgroup. Figure 6.3 highlights a flowchart for the ancestral genome content reconstruction. A variety of in-house perl scripts were written for the post-processing of BLASTP outputs and are available upon request. Species-specific gene lists with reciprocal hit scores were generated. A reciprocal hit scores could be 0, 1, and 2; 0 indicates that one species has a gene that is not present in the other species (which may be caused by gene gain/loss); 1 indicates the presence of homologous sequences in both species but they are not BBH (Bidirectional/Reciprocal Best Hit) (which may be caused by gene duplications); 2 indicates the presence of homologous sequence in both species and these two sequences are BBH. These scores were then used to summarize the presence/absence for each gene in each species.

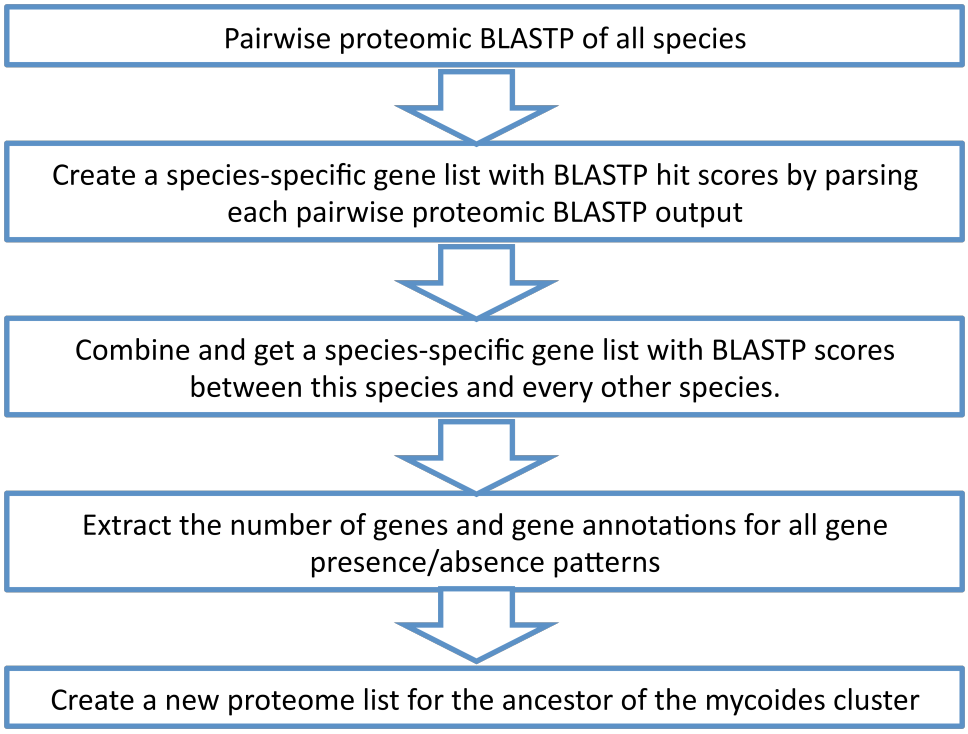


Figure 6.3: The flowchart for genome content reconstruction.
Each gene can be either present or absent in each of the five species, so there are thus a totally 32 combinations (2^5) for a particular gene's status of presence/absence in the five species, obviously the combination of a gene being absent in all five species is not considered. Therefore 31 scenarios were considered, and the number of genes and their annotations for each scenario were summarized. The presence/absence of a gene in the ancestor is determined by minimizing the number of changes (the number of gene gains or losses) from the ancestor to the extant species. The inferred proteins were used to compile a proteomic list for the genome content of the last common ancestor of mycoides (mycoides cluster ancestor - MCA).

Comparing genome content of the mycoides cluster ancestor (MCA) with two hypothetical minimal genomes of mycoplasma

Figure 6.5 provides a flowchart of the procedures used to compare genome content. First, the proteomic list representing the genome content of the MCA was compared with the proteomic lists of the hypothetical minimal genomes of *M. genitalium* containing 381 genes (Mg381) [146] and *M. pulmonis* containing 310 genes (Mp310) [147] via pair-wise BLASTP with an e-value cutoff 0.01. Genes present in the MCA genome but not present in either Mg381 or Mp310 were individually extracted and annotated as MCA_absentMg381 and MCA_absentMp310, respectively. Next, reciprocal BLASTPs

with an e-value cutoff 0.01 were performed to compare gene content between MCA_absentMg381 and the whole proteome of *M. genitalium* that contains 476 genes (Mg476) to determine genes present in both the ancestral and modern genomes but that are absent from the minimal genome based on genes with reciprocal best hits. The same analyses were performed between MCA_absentMp310 and the whole proteome of *M. pulmonis* that contains 782 genes (Mp782). Genes matching these criteria were extracted and annotated as MCA_absentMg381_presentMg476 and MCA_absentMp310_presentMp782. Lastly, individual genes represented in both of these two latter lists were extracted. The extracted genes correspond to those genes present in the three whole genomes of MCA, Mp and Mg, but not in the minimal genomes of Mp and Mg.

Functional annotations of gene lists:

KAAS (KEGG Automatic Annotation Server: www.genome.jp/kegg/kaas/) [160] was used to find the pathways associated with specific genes using protein sequences as input.

Ancestral genome rearrangement reconstruction:

Locally collinear blocks (LCBs)/syntenies were generated by progressiveMAUVE [161] implemented in barphlye-0.0.0 [162, 163] and used as input for genome rearrangement reconstruction using MGR-2.03 [164] based on the maximum parsimony algorithm. The genomic locations of genome rearrangement breakpoints were extracted and mapped to the whole genome annotation of the mycoides cluster, and visualized using NCBI Genome Workbench 2.4.5 (http://www.ncbi.nlm.nih.gov/tools/gbench/).

Results and discussion

Mycoplasma species tree and the search of outgroup

Our attempt to reconstruct an ancestral genome of the mycoides cluster containing four species (MSC, MLC, MCAP, and MSB) required that we first identify an outgroup in order to root the cluster at its last common ancestor. The search to find a closely-related outgroup to the mycoides cluster did not reveal a suitable species within the taxonomic group of *Mycoplasma* based on our phylogenetic analyses of Mycoplasma-annotated species. A literature search suggested that Mollicute taxonomic names do not necessarily reflect evolutionary relationships, and particularly, the mycoides cluster appeared to be more similar to *Mesoplasma florum* (Mf) than to other Mycoplasma species [165]. We therefore conducted four independent evolutionary analyses to resolve the annotation confusion of *Mesoplasma*/*Mycoplasma* (Figure 6.1). All four analyses (16S rRNA, presence/absence of COGs, syntenies, and averaged distances between species based on ortholog comparisons) support the position of *Mesoplasma* as the closest relative to the mycoides cluster instead of any other Mollicute not from the mycoides cluster including species in the *Mycoplasma* genus, where the mycoides cluster belong taxonomically (Figure 6.1). The monophyletic group containing *Mesoplasma* and the mycoides cluster was statistically supported by having a posterior probability of 1.0 in the MrBayes analysis and bootstrap values of 100% in both parsimony and neighbor-joining algorithms in PAUP analyses for both 16S rRNA and presence/absence of COGs. The closer relationship between the mycoides cluster and *Mesoplasma*, as opposed to other Mycoplasma species, highlights short-comings associated with some taxonomy classifications, and emphasizes the importance of choosing evolutionarily distinctive phenotypes as well as integrating molecular data for biological classifications/nomenclatures [166].

Mesoplasma (NCBI accession NC_006055) was therefore used as the outgroup for the ancestral genome reconstruction of the mycoides cluster. The outgroup *Mesoplasma* is biologically distinct from the mycoides cluster in that the former use plants and insects as hosts instead of ruminants for the latter. *Mesoplasma* also lack pathogenicity compared to the mycoides cluster.

Ancestral genome content reconstruction

Whole proteome pair-wise BLASTP was performed on the mycoides cluster (MLC, MSC, MCAP and MSB) and their outgroup *Mesoplasma florum* (Mf). The number of genes was summarized for all 31 scenarios (gene presence/absence patterns discussed in the Methods section) in Figure 6.4A. Overall, the number of genes classified into the different scenarios varied substantially, with 466 genes present in all five species. The number of species-specific gene losses (varying from 3 to 16) is much smaller than the number of species-specific gene gains (from 53 to 245). The only exception to this observation is the outgroup Mf lineage, which has 161 unique gene losses and 177 unique gene gains compared to the ingroup mycoides cluster.

The gene content of the mycoides cluster ancestor (MCA) was reconstructed for 20 out of the total 31 scenarios. Twelve of these 20 scenarios predict the presence of genes in the MCA ancestor assuming the fewest number of gene losses/gains throughout the tree that best explain presence/absence of genes among modern species. The other eight scenarios all require either two gene gains or losses in the MCA genome and are thus equally likely and require additional analyses to resolve the ambiguity (discussed below). The remaining 11 of the 31 scenarios were not considered because they failed to minimize the number of gains/losses on the tree to explain the data.

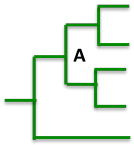
Figure 6.4B highlights the 20 scenarios (patterns P1 to P20) and the number of genes under each of the scenarios. The eight unresolved scenarios from above are identified as patterns P6, P7, P11, P12, P16, P17, P19 and P20 and contain a total of 88 genes. These 88 genes were further analyzed for their presence/absence in the MCA ancestor by performing BLASTP against the whole proteomes of two more-distantly related species *Aster yellows witches'-broom phytoplasma AYWB* (AYWB, NCBI accession NC_007716) and *Onion yellow phytoplasma OY-M* (PAM, NCBI accession NC_005303). Based on our criteria, between 4 and 6 out of the 88 genes were determined to be present in the ancestor. Two genes could not be resolved because they are only present in one of the two *Phytoplasma* species. Thus, these two genes may have been either gained or lost three times during evolution. We elected to include the two genes (MCA_667 and MCA_668) in the MCA ancestor proteome list but labeled them with asterisks. Overall, 668 genes were identified as being present in the MCA ancestor, with a majority (627) coming from the combination of 466 core genes conserved in all five species (P1) and 161 genes unique just to the mycoides cluster (P2).

All 668 genes identified in the MCA ancestor were annotated by genes from the mycoides cluster, particularly by the representative species MLC, and supplemented by MCAP (Supplementary table E.1). Of the modern mycoides species, the MLC genome contains the most genes that are inferred to also be present in the MCA ancestral genome. For genes inferred in the MCA genome but not present in the MLC genome, the MCAP genome was used to annotate the ancestral genes (only five such cases; P8, P10 and P11). Thus 663 of 668 genes in the MCA ancestor were annotated with MLC locus tags, while

the other five genes (MCA_640 to MCA_643, and MCA_663) were annotated with MCAP locus tags (Supplementary table E.1).

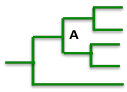
Functions for the 161 genes present in the mycoides cluster but absent from the *Mesoplasma* outgroup were assigned using KAAS and are implicated in the divergence between mycoides and *Mesoplasma*. Many of these genes are annotated as lipoproteins, ABC transporters, transposases, and other genes involved in pathways of membrane transport and carbon metabolisms such as the phosphotransferase system (PTS), glycerol import and metabolisms, mannitol uptake and metabolisms, and amino sugar (*N*-Acetylmuramic acid - MurNAc and *N*-Acetylglucosamine - GlcNAc) metabolisms (Figure 6.4C). These include genes that have been reported to be responsible for bacterial pathogenicity such as ABC transporters [167], lipoproteins that help evade the host immune systems [168], transposases that increase genome plasticity and bacterial pathogenicity via acquisition of antibiotic resistance [169], as well as genes involved in glycerol import/metabolism that harm host cells by creating hydrogen peroxide as a byproduct [170]. These also include other import and metabolism genes that may be responsible for carbon-source utilization and this in turn can determine host specificity, such as ruminant hosts for mycoides and plant/insect hosts for *Mesoplasma* [171]. In summary, many of the 161 genes common among mycoides, but absent from the outgroup *Mesoplasma*, appear to be responsible for both the pathogenicity of the mycoides cluster (but that is absent from *Mesoplasma*) and the ruminant host-specificity of this cluster.

A:



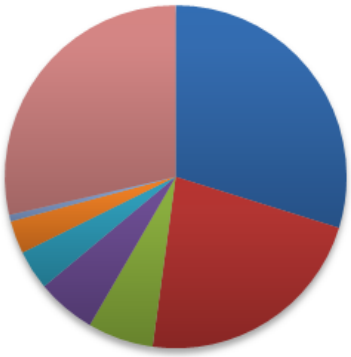
Species	Gene Gains and Loss																														
MSB	+	-	+	+	+	+	+	-	-	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	+	+	+	+	-	-	-
MCAP	+	+	-	+	+	+	-	+	-	-	-	-	+	+	-	+	+	-	-	-	-	+	+	+	-	+	+	-	+	-	-
MSC	+	+	+	-	+	+	-	-	+	-	-	+	-	-	+	-	+	-	+	+	-	+	-	+	+	-	+	-	-	+	-
MLC	+	+	+	+	-	+	-	-	-	+	-	+	-	-	+	+	-	+	-	-	+	-	+	+	+	+	-	-	-	-	+
Mf	+	+	+	+	+	-	-	-	-	-	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+
Counts	466	3	16	8	3	161	85	53	245	119	177	1	1	28	30	1	0	2	0	31	21	11	27	11	31	23	12	0	4	5	2

B:



	Gene Presence/Absence Patterns																			
Patterns	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
MSB	+	+	+	-	-	-	+	+	-	+	-	+	+	-	+	-	+	+	-	+
MCAP	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
MSC	+	+	-	-	+	+	-	+	+	-	-	+	+	+	-	-	+	+	+	-
MLC	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	+	+	-	-	-
Mf	+	-	+	+	+	-	-	+	+	+	+	-	+	+	+	+	-	+	+	+
Counts*	466	161	8	1	3	1/1 1	1/2 3	3	0	1	1/4	0/1 2	16	1	2	0/2	1-3/3 1	0	0/5	0
ANC Presence	+	+	+	+	+	??	??	+	+	+	??	??	+	+	+	??	??	+	??	??
ANC Counts	627		12			2		4			1		19			1-3?		0		
ANC Total	666 - 668?																			

C:



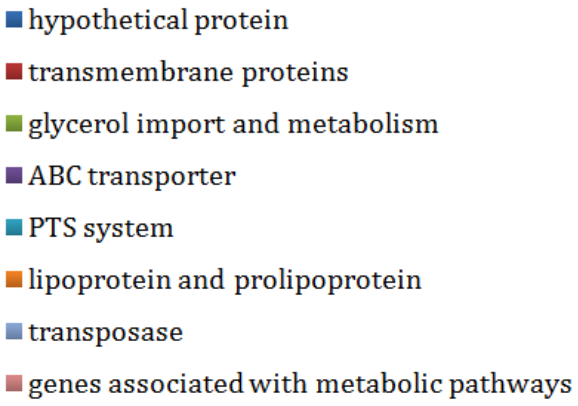


Figure 6.4: Genome Content Reconstruction and functional distribution of genes unique in the mycoides cluster.

A: Summary of 31 patterns of gene presence/absence. Red color highlights gene presence and purple color highlights gene absence. '+/-' represents the gene presence/absence.

B: Summary of 20 out of 31 patterns that may contain genes in the MCA ancestor based on the parsimony algorithm. '??' represents unresolved patterns that were further analyzed by adding two *Phytoplasma* species. '*' contains the numbers colored in red in the format of 'a/b', with 'b' indicating the count of genes in the particular pattern, and 'a' representing the count of genes in the MCA ancestor after considering two *Phytoplasma* species. '1-3?' is due to the two unresolved genes in the MCA ancestor, and the total number of genes in the ancestor is from 666 to 668.

C: Functional distributions and metabolic pathways for 161 gene present in the mycoides cluster but absent in the outgroup *Mesoplasma*.

Comparing genome content of the mycoides cluster ancestor (MCA) with two hypothetical minimal genomes of mycoplasma

Figure 6.5 shows a flowchart of the procedures and results comparing the genome content of the ancestral MCA genome to the hypothetical minimal genomes of two different mycoplasmas**.** A BLASTP comparison of gene content between the MCA ancestor and the minimal genome of *M. genitalium* (Mg381) identified 345 genes present in MCA but absent from Mg381 (MCA_absentMg381). A similar comparison between the MCA ancestor and the minimal genome of *M. pulmonis* (Mp310) identified 355 genes present in MCA but absent from Mp310 (MCA_absentMp310). A reciprocal BLASTP comparisons between the 345 MCA_absentMg381 genes and the whole proteome of *M. genitalium* (Mg476) identified 41 genes present in the ancestral MCA genome and the whole genome of *M. genitalium* but absent from the minimal *M. genitalium* genome (MCA_absentMg381_presentMg476). These 41 genes are bidirectional/reciprocal best hits (BBH) and were labeled according to the gene identifier lists of *M. genitalium* and the MCA (Supplementary table E.2). Similarly, a reciprocal BLASTP comparison between the 355 MCA_absentMp310 genes and the whole proteome of *M. pulmonis* (Mp782) identified 105 genes present in the ancestral MCA genome and the whole genome of *M. pulmonis* but absent from the minimal *M. pulmonis* genome (MCA_absentMp310_presentMp782) (Supplementary table E.3).

Combining these genes to their respective minimal genomes increases the number of genes we predict to be minimally required to sustain their free-living life cycles. Thus, the essential gene sets for Mg and Mp should be expanded to '381+41=422' and '310+105=415' respectively. We suspect that the design of the Mg and Mp gene-disruption experiments were designed to select for genes necessary for short-term viability but failed to identify many genes necessary for long-term survivability. The list

of 41 and 105 genes displays an abundance of genes involved in metabolism and DNA repair systems (Supplementary table E.2 and E.3). A comparison of the 41 MCA_absentMg381_presentMg476 genes and the 105 MCA_absentMp310_presentMp782 genes shows that 23 genes overlap (Supplementary table E.4). Many of these 23 genes support metabolic processes. The dispensability of metabolic genes in the gene-disruption experiments is probably due to the rich culture media containing supplements used in these experiments [146, 147]. Alternatively, three of the 23 genes are involved in DNA repair (MG_339/MYPU_2520 *recA*, MG_352/MYPU_0860 *recU*, and MG_360/MYPU_1880 *mucB*) (Supplementary table E.4). Again, the dispensability of DNA repair genes in the gene-disruption experiments is probably due to the fact that only a single generation was used to determine viability post disruption. The inability to repair mutations in the genome will, of course, become exacerbated over time as random mutations occur with each passing generation. Experimental evidence supports this notion even in *M. pulmonis*. Cells lacking DNA repair genes MYPU_2520 (*recA*), MYPU_7100 (*uvrA*), MYPU_0960 (*uvrB*), and MYPU_1560 (*uvrC*) are highly susceptible to UV irradiation [147]. With the exception of *uvrA*, these DNA repair genes are present in the list of 23 genes discussed above (Supplementary table E.4).

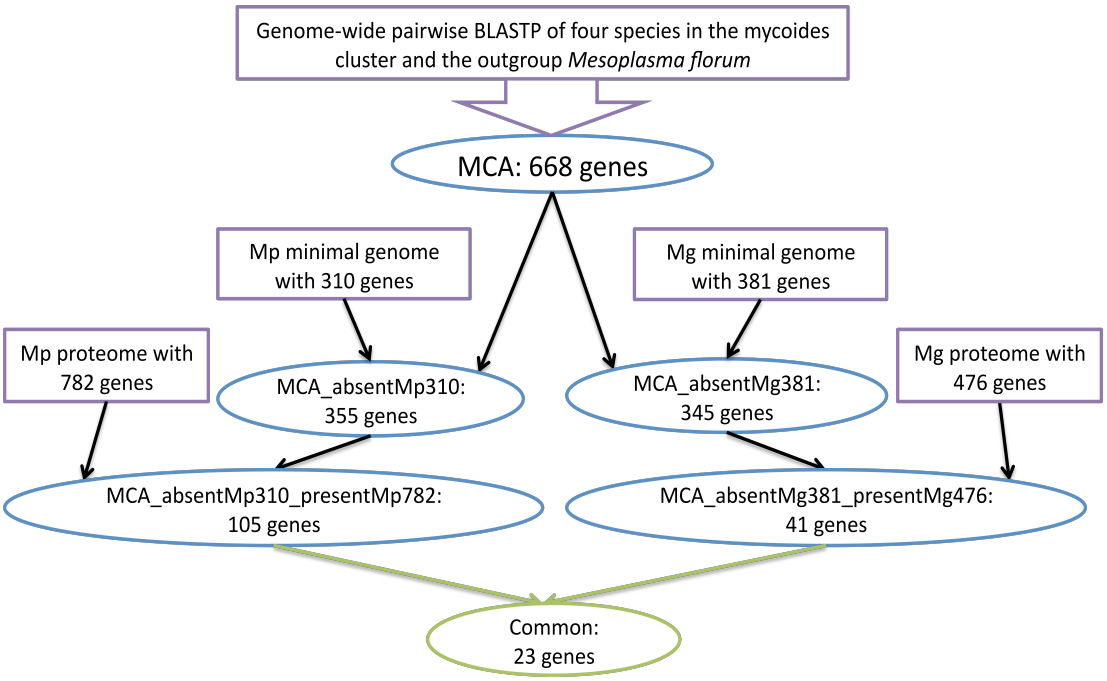


Figure 6.5: Flowcharts of the procedures and results for genome content comparison of the mycoides cluster ancestor (MCA) and two hypothetical minimal genomes of *M. genitalium* (Mg) and *M. pulmonis* (Mp).

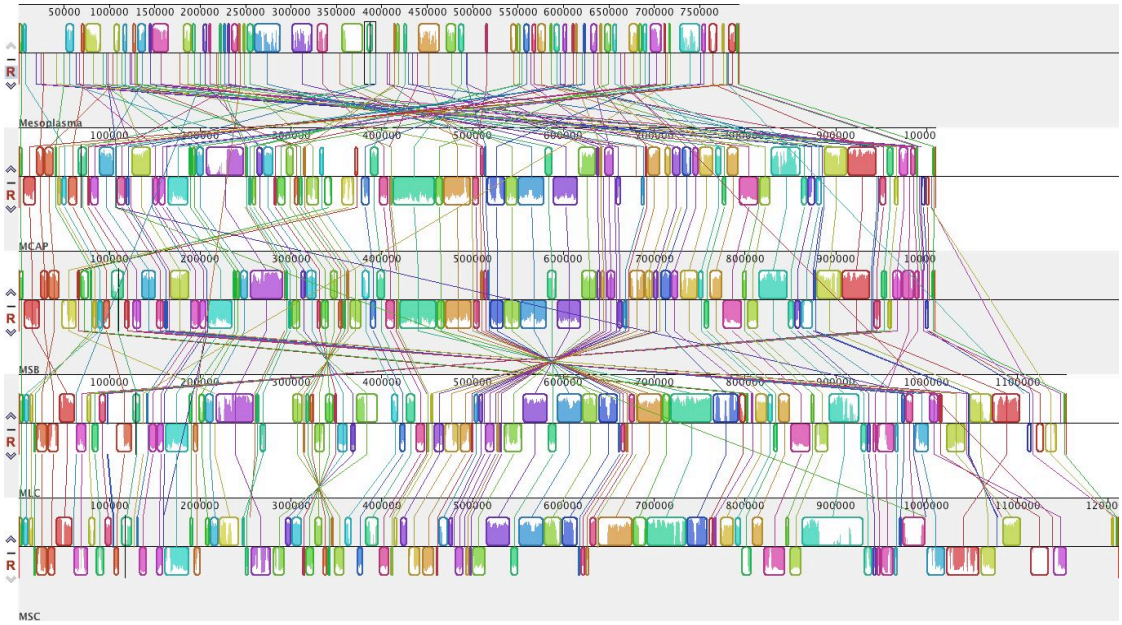
Moreover, the method itself of using transposon insertions may also lead to an underestimation of essential genes. For *M. pulmonis*, only 321 genes were inactivated in the proteome of 782 genes, which means 461 genes were not inactivated and thus should be considered essential, however, only 310 genes were identified as essential [147]. From that study [147], 151 (461 minus 310) genes were ruled out due to the length of these genes being shorter than 1000 base pairs (bp) and they were not contained within an operon – the method of gene interruption using transposons is not dense enough to cover all genes shorter than 1000bp. Of these 151 genes, we identified 28 genes contained within the 105 MCA_absentMp310_presentMp782 genes, and all of the 28 genes are shorter than 1000bp except one, MYPU_1880 (*mucB*). We suspect that the lack of coverage density using transposon insertions causes an underestimation of essential genes

in *M. pulmonis* and that including these 28 genes, along with the remaining 105 genes, into a minimal genome is more likely to generate a self-sustaining organism.

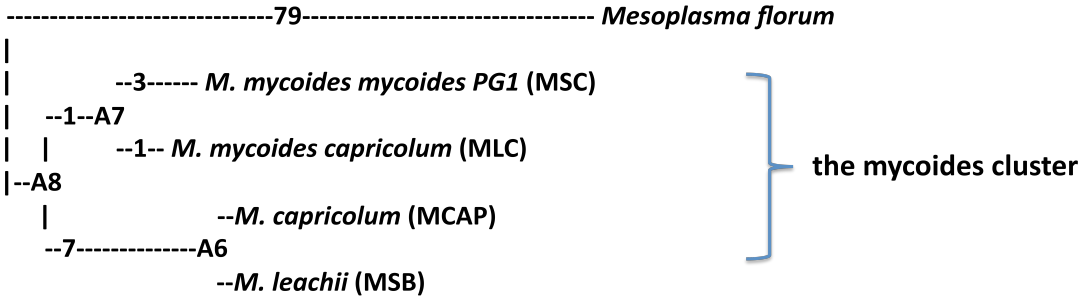
Genome rearrangement reconstruction

In addition to gene content, large-scale reconstruction of an ancestral genome requires that we infer genome rearrangements from the ancestral genome to its modern descendents. A multiple genome alignment of the mycoides cluster and the outgroup *Mesoplasma* by progressiveMAUVE generated 93 conserved Locally Collinear Blocks (LCBs)/syntenies (Figure 6.6 A). The arrangements of these syntenies were reconstructed for the MCA ancestor assuming the minimal number of permutations from the ancestor to any of the modern genomes (Figure 6.6 B). The NCBI Genome Workbench was used to visualize the genome rearrangement breakpoints. The total number of synteny rearrangements is inferred to be 91 across the whole tree; 1 from A8 (the MCA ancestor) to A7 (the ancestor of MLC and MSC), 7 from A8 to A6 (the ancestor of MCAP and MSB), 79 from A8 to the outgroup *Mesoplasma*, 3 from A7 to MSC, 1 from A7 to MLC, and 0 from A6 to MCAP/MSB (MCAP and MSB have exactly the same order of syntenies). Rearrangement breakpoints occurring from A7 to MSC/MLC were found close to or within endogenous transposases, but such features were not found for rearrangement breakpoints occurring from A8 to A6 in MCAP/MSB. This suggests that other genome rearrangement mechanisms beyond transposition occur within some lineages. Further understanding of genome rearrangement mechanisms can help us understand the evolution of genome structure and bacterial pathogenicity [169].

A:



B:



>A8
1 2 91 -7 -16 -5 92 -4 6 -28 65 -3 -51 -11 44 45 -10 -89 -88 -87 13 86 19 70 -31 33 -62 -61 29 8 -71 -81 49 -21 -18 48 -37 -63 50 -41 53 -32 -46 -76 -60 -75 34 35 -64 -74 -58 -72 -52 -55 40 -38 39 43 83 -57 -47 -42 59 -22 54 77 56 79 80 82 -23 66 -26 67 -27 78 -68 -25 -30 -20 -90 -12 -69 9 15 17 -85 -84 24 -36 -73 14 93

Figure 6.6: Multiple genome alignment and genome rearrangement reconstruction of the mycoides cluster.

A: Multiple genome alignment of the mycoides cluster and the outgroup *Mesoplasma* with 93 Local Collinear Blocks (LCBs)/syntenies.

B: Ancestral genome reconstruction of the mycoides cluster. A8 is the ancestor for the mycoides cluster, and its syntenies were listed. The order of the number represents the order of syntenies, and the sign of the number represents the orientation of syntenies.

The 93 LCB/syntenies contain 466 of the 668 genes we have inferred for the mycoides ancestral genome. The order of genes within the syntenies will be necessary for future studies intended to synthesize the ancestral genome in the laboratory. 161 of the

remaining 202 genes are conserved within mycoides but absent from the outgroup but their gene orders can be inferred from a comparison among mycoide genomes only.

Conclusion

We have for the first time extended ancestral reconstruction from the single-to-a-few gene level up to the whole genome level. Both genome content and genome rearrangements were reconstructed for Mycoplasma species. The ancestral genome reconstruction (AGR) helps us understand evolution on the genomic level, including the association of genotypes and phenotypes. For instance, our study identified genes that are likely responsible for pathogenicity and host-specificity among Mycoplasma species. We also identified metabolic and DNA-repair genes in our ancestral genome that are absent from other minimal genome studies [146, 147]. This result is not surprising given that experimental knockout studies focus on single-generation survival [146, 147], thus genes required for long-term survival are not identified.

The field of synthetic biology seeks a minimal genome sufficient to support a viable, self-sustaining organism. This genome is considered to be a chassis upon which additional genes can be added to support whatever desired functionality researchers require of a particular organism. We have attempted to exploit evolutionary analyses to better understand the concept of a minimal genome and how this compares to an ancestral genome. Minimal genomes have been inferred using computational and experimental techniques. On the computational side, a minimal genome is inferred by comparing modern genomes and then selecting all the genes in common among the genomes. Such comparisons provide insight into core genes but these studies can be limited by functional divergence among the different modern organisms that would prevent us from identifying

all of the genes essential for producing a recombinant minimal organism. On the experimental side, current studies have not exhaustively targeted all genes in Mycoplasma.

Our inference of gene content for an ancestral Mycoplasma organism is substantially different than the inferred minimal gene content for Mycoplasma organisms [146, 147]. We anticipate that our ancestral gene content will be exploited in two directions. In one direction, this information will be used to synthesize and resurrect an ancient genome in the laboratory using recent genome synthesis and transplantation technology [132-134]. In another direction, this information can be used to determine which genes should and should not be removed from a modern genome being used to generate a minimal genome.

In total, our study provides insight into the evolution of Mycoplasma species and also provides a list of genes to the synthetic biology community that can be utilized to create an ancient or engineered organism.

Abbreviation

ASR, Ancestral Sequence Reconstruction; AGR, Ancestral Genome Reconstruction; Trp, Tryptophan; TBR, tree-bisection-reconnection; MCA, Mycoides Cluster Ancestor; MSC, *M. mycoides mycoides*; MLC, *M. mycoides capricolum*; MCAP, *M. capricolum*; MSB, *M. leachii*; Mf, *Mesoplasma florum*; Mp, *M. pulmonis*; Mg, *M. genitalium*. AYWB, *Phytoplasma AYWB*; PAM, *Phytoplasma OY-M*; bp, base pair; COG, Cluster of Orthologous Groups; KAAS, KEGG Automatic Annotation Server; PTS, Phosphotransferase system; BBH, Bidirectional Best Hit.

Acknowledgements

This work was supported by Georgia Institute of Technology and NASA to Eric A. Gaucher. We thank to the technical support from ‘The Partnership for an Advanced Computing Environment (PACE)’ at Georgia Institute of Technology.

CHAPTER 7

COMPUTATIONAL VALIDATIONS OF ANCESTRAL SEQUENCE RECONSTRUCTION

Abstract

Ancestral sequence reconstruction (ASR) has been wildly applied to recover ancient molecules and study their biochemical functions to help understand origin and evolution of life. Consequently, the correctness and accuracy of computational ASR is essential for subsequent experimental investigations to answer biological questions. We performed computational simulations to validate ASR of Hobbs' resurrection of 3-isopropylmalate dehydrogenase (LeuB) of *Bacillus* and inference of thermophily evolution. We failed to reveal a high accuracy and correctness of ASR based on Hobbs' LeuB tree topology. Our phylogenetic inference further confirmed our concerns about the uncertainty of the LeuB tree topology. Besides, we failed to repeat Hobbs' ASR using various evolutionary models and datasets. Overall, our studies of LeuB phylogeny, ASR and computational simulations indicate that an alternative evolutionary pattern of thermophily revealed by LeuB in *Bacillus* might be achieved. However, experimental validation of our ASR and a confident LeuB phylogeny are required.

Introduction

Computational ancestral sequence reconstruction (ASR) has been utilized to infer genetic codes in ancient molecules, followed by experimental resurrection and biochemical studies of ancient genes, to help understand origin and evolution of life, as

well as inference of changes of climate, ecology and physiology [34-37]. Thus accuracy and correctness of computational ASR is essential for subsequent experimental investigations to understand origin and evolution of life. Accuracy herein refers to the probability of an inferred amino acid compared with the other 19 amino acids for a particular site of a particular ancestral node, and overall accuracy refers to the averaged accuracy over all sites for a particular ancestral node. Whereas, correctness refers to the percentage of identity by comparing the computationally inferred ancestral sequence with the true ancestral sequence (true ancestral sequences can be from computational simulation or experimental evolution).

Recently, Hobbs and etc. used ASR to resurrect 3-isopropylmalate dehydrogenase (LeuB) of *Bacillus*, and proposed that thermophily emerged multiple times as demonstrated by thermostability of a series of ancient LeuB genes in *Bacillus* [172]. The evolutionary pattern of thermophily revealed by Hobbs contradict our ancestral resurrection using EF-Tu and thioredoxin, which indicates that thermophily was a primitive trait and it experienced a gradually loss along evolution [38, 39]. Meanwhile, we are concerned about the correctness and accuracy of LeuB ASR by Hobbs, since the phylogenetic tree of LeuB in *Bacillus* used by Hobbs for ASR is not statistically supported, particularly two branches having two of four resurrected ancient genes (ANC2 and ANC3) do not have bootstrap supports (Figure 1 in Hobbs' paper). Additionally, LG model was the selected best model for amino acid sequences, while JTT model was used in ASR by Hobbs, and Hobbs emphasized one ASR criteria for determining an amino acid under bias is choosing the amino acid not affecting thermophily, however, no such patterns is available associating 20 amino acids with thermophilic status.

Therefore, we repeated Hobbs' phylogeny reconstruction using both maximum likelihood and Bayesian algorithms, ancestral sequence reconstruction of LeuB using various models and datasets including DNA, codon and amino acid. We also performed computational simulations for the accuracy and correctness of ASR given Hobbs' tree. Our phylogenetic analyses of LeuB confirmed our concern about the uncertainty of the tree topology; Ancient LeuB sequences inferred by us differed substantially from Hobbs' ASR, and our computational simulations did not real high accuracy and correctness of ASR by using Hobbs' LeuB tree. Overall, our results from LeuB phylogeny, ASR and simulations indicate that different evolutionary patterns of thermophily might be achieved given a confident tree topology and validated ASR. However, subsequent experimental validation of our computationally referred ancestral LeuB genes is required. Additionally, LeuB phylogeny with a higher resolution and confidence is required by having more sequences available from diverse *Bacillus* species.

Methods

Phylogeny of LeuB in *Bacillus*

LeuB protein and DNA sequences in *Bacillus* and *Clostridium* were extracted based on the accession numbers provided in the supplementary material table S1 of Hobbs' paper. Multiple sequences alignment was performed using ClustalW2 [90] and MUSLCE [63], with manual modifications. Protein model was selected for LeuB using ProtTest 2.4 [4]. A maximum likelihood tree of LeuB proteins was constructed using Garli 1.0 [173] with LG+I+G model and 1024 bootstrap replicates. Bootstrap replicates were summarized with bootstrap values mapped to the LeuB tree topology used by Hobbs with SumTrees implemented in DendroPy-3.10.1 [174]. A Bayesian tree of LeuB proteins was reconstructed using MrBayes [79], with WAG+I+G model, two independent runs, four

chains in each run and four million generations. The standard deviation of split frequency is smaller than 0.005. Phylogenetic trees labeled with bootstrap values from Garli and posterior probability from Mrbayes were visualized using FigTree v.1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

Computational Simulation of ancestral sequence reconstruction

Protein sequences were simulated with 100 replicates by evolver in PAML44 [98], based on the same protein length, the same number of taxa, and the same tree topology of LeuB used by Hobbs'. Protein parameters used by evolver were predicted by ProtTest, though LG+G model was used instead of LG+I+G model, because the proportion of invariable sites (I) is not implemented in PAML. Ancestral sequences of simulated protein sequences were reconstructed for all 100 replicates using codeml in PAML44. Ancestral sequences and overall accuracies were extracted from PAML outputs for 100 replicates individually for interested ancestral nodes including ANC1 to ANC4 in the Hobbs' paper. Ancestral sequences reconstructed by PAML were compared with the true ancestral sequences generated during simulation, and the correctness of ancestral sequence reconstruction was calculated and averaged for all 100 replicates.

Ancestral sequence reconstruction of LeuB

Ancestral sequence reconstruction of LeuB in *Bacillus* was performed using PAML44. DNA sequence reconstruction was using baseml with GTR/REV model, and codon and protein sequence reconstruction was using codeml. For the codon sequences, aaDist from 0 to 6 were implemented independently, and aaDist considers physicochemical properties of different amino acids and relative substitution frequency. Codon models includes geometric relationship using Grantham (aaDist=1), Miyata &

Yasunaga (aaDist=2), 1974c - composition (aaDist=3), 1974p - polarity (aaDist=4), 1974v - volume (aaDist=5), and aromaticity (aaDist=6). Grantham combines three factors, composition (the ratio of non-carbon to carbon in the side chain), polarity, and volume, and is a more complicated model than 1974c (aaDist=3), 1974p (aaDist=4) and 1974v (aaDist=5) individually by considering only one of the three factors [175]. For protein sequence reconstruction, JTT and LG models were implemented respectively, combined with empirical or empirical+F, and estimated alpha or fixed alpha 0.58 (the value is from ProtTest) independently. The reconstructed LeuB ancestral sequences using different datasets and implementing different models were compared with those in Hobbs' paper. MEGA 5 [176] was used to compute pairwise distance of computationally reconstructed LeuB by us and Hobbs. Different inferred sites were extracted and mapped to the structure of ANC4 (PDB file '3U1H') and visualized using PyMOL 1.3 [177].

Results and discussion

Validation of LeuB phylogenetic tree used in Hobbs' paper

The tree topology of LeuB utilized in Hobbs' paper was validated by using both bootstrap supported Garli (LG+I+G model) with a maximum likelihood algorithm and posterior probability supported MrBayes (WAG+I+G model) with a Bayesian algorithm. Protein model LG+I+G was selected for LeuB proteins using ProtTest, and it is consistent with Hobbs' model selection. Since LG model is not implemented in MrBayes, a mixed amino acid model was initially used and WAG model was selected and fixed as shown in MrBayes parameter outputs, thus WAG model was finally used to rerun MrBayes. While, JTTmodel was used in MrBayes by Hobbs and we don't know why it was used [172].

The bootstrap and posterior probability values were labeled for ancestral nodes from ANC1 to ANC4 (Figure 7.1). As expected, the bootstrap supports for ANC2 and ANC3 are too small to confidently support the branch, based on a bootstrap value of 0.7 corresponding to 95% confidence [7]. However, the corresponding posterior probabilities are 0.87 and 0.92, in contrast to bootstrap values 0.23 and 0.42 for ANC2 and ANC3 (Figure 7.1). It indicates an overconfidence of posterior probability by MrBayes compared with bootstrap statistics [178]. Individual trees by Garli and MrBayes with branch lengths and branch supports were provided (Supplementary figure F.1).

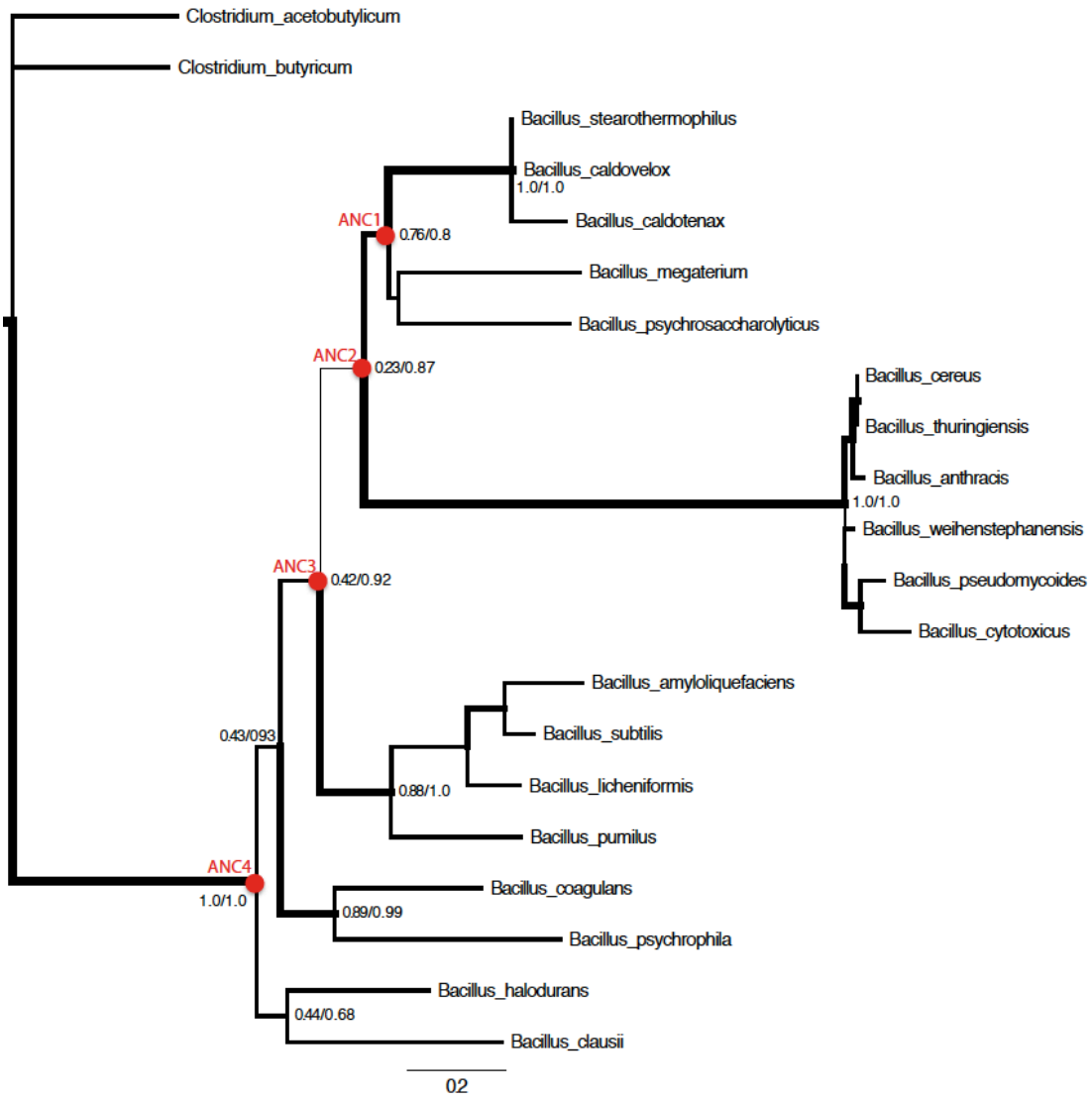


Figure 7.1: The validation of the tree topology of LeuB utilized by Hobbs' with branches supported with bootstrap values from Garli and posterior probability values from MrBayes.
The branch length is based on Garli, and interested internal nodes including ANC1 to ANC4 were labeled with bootstrap values on the left side of the slash and posterior probability on the right side of the slash. The thickness of branches is based on bootstrap values from Garli.

Overall, our phylogenetic analyses of LeuB did not support branches containing ANC2 and ANC3, which means the branch leading from ANC1 may swap to other positions in the tree. Species tree of *Bacillus* were not consistent for the position of branch containing ANC1 by using different datasets (genomic data and 16S rRNA) and different methods (NJ and ML), and the support values are low [179]. In this case, the

ancestral nodes ANC1 to ANC4 may be swapped and the ancestral sequences will be different based on different tree topologies. Even with the same ASR, the explanation of thermophily evolution will change, since ANC1 containing clade is the thermophilic clade, if the clade is closer to the root, then thermophily is not stochastic, but a primitive trait with gradually loss along evolution. Therefore, we performed computational simulations to check the correctness and accuracy of ASR based on this specific tree topology.

Computational Simulation of ancestral sequence reconstruction

The accuracy and correctness of LeuB ancestral sequence reconstruction using the tree topology by Hobbs were examined. Protein sequences were simulated 100 times and ancestral sequences were reconstructed for each of the 100 replicates individually. Accuracy and correctness of ancestral sequence reconstruction were retrieved and plotted for 100 replicates (Supplementary figure F.2). Averaged accuracy and correctness of ancestral sequence reconstruction for Clade1/ANC1, Clade2, Clade3, ANC2, ANC3, and ANC4 are 0.866, 1, 0.915, 0.876, 0.874, and 0.888 (accuracy), and 0.896, 0.995, 0.912, 0.906, 0.897, and 0.913 (correctness) (Figure 7.2). Overall, none of the four ancestral nodes from ANC1 to ANC4 have high accuracy or correctness, with values around 0.9. The only exception is Clade 2 with almost 100% accuracy and correctness, and a long internal branch leading to it and statistically supported short external branches. The topology and features of branch length for Clade 2 (the branch containing *B. anthracis*) can be better visualized in Figure 7.1. In contrast, other interested branches have long external branches and short internal branches, and lower accuracy and correctness in ancestral sequence inference. Such effects (the association of low accuracy and

correctness in inferring ancestral sequence with short internal and long external branches) have been simulated and revealed by Joseph Thorton's group [44].

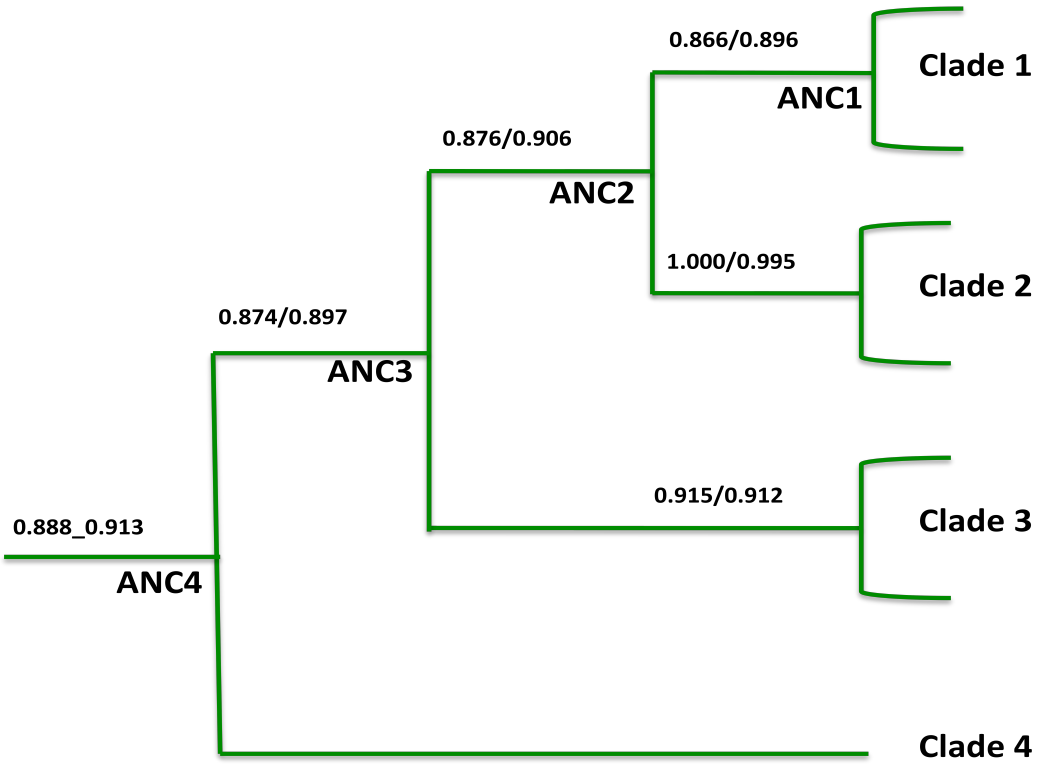


Figure 7.2: Accuracy and correctness of ASR by computational simulations.
The tree topology and four labeled ancestral nodes from ANC1 to ANC4 are the exactly the same as those in Hobbs' paper [172], with four clades to represent species for short. Amino acids were simulated with 100 replicates by PAML evolver and ASR was performed by PAML codeml. Overall accuracy by PAML reconstructed for interested nodes are averaged for 100 replicated and labeled on the left side of the slash; Correctness by comparing PAML reconstructed sequences and the true ancestral sequences from the simulated data are averaged for 100 replicates and are labeled on the right side of the slash.

Therefore, the specific evolutionary patterns possessed by LeuB in *Bacillus* may cause low accuracies and correctness in ancestral state reference for all four interested ancestral nodes from ANC1 to ANC4, and the uncertain ancestral sequences might lead to ambiguous phenotypes including thermophily.

Ancestral sequence reconstruction of LeuB

We redid ASR for LeuB by using the same sequences and tree topology as Hobbs, and we compared our reconstructed ancestral sequences with Hobbs. In general, none of our ancestral sequences of LeuB inferred using different models and datasets (DNA, codon and amino acid) were the same as Hobbs'. For protein sequence reconstruction of LeuB, we did not detect substantial difference in the comparison by using empirical or empirical+F, and estimated alpha or fixed alpha using JTT model, or using different versions of PAML (PAML43 and PAML44). Multiple sequence alignment of ASR with various models for DNA, amino acid and codon were compared to each of Hobbs' ANC from ANC1 to ANC4 individually (Supplementary figure F.3). We only selected JTT, WAG and LG models to represent protein sequence reconstructions, since other models did not make closer ASR to Hobbs'. JTT is the model used by Hobbs, LG is the model selected by ProtTest, though we don't know why Hobbs did not use LG model, and WAG model is the second best model except LG for LeuB sequences and WAG was implemented in MrBayes for phylogenetic inference. For codon models, we only selected codon models Grantham (aaDist=1) and Miyata & Yasunaga (aaDist=2) as representatives, and other models did not make closer ASR to Hobbs'. MEGA pairwise distance comparison was performed to show the number of different sites between our ASRs under representative models and Hobbs' ASRs. Particularly for ANC1, my ASR differed from Hobbs' with 21 (JTT model) to 39 (DNA model) amino acids; For ANC2, different sites are from 25 (aaDist2 model) to 34 (LG model); For ANC3, different sites are from 25 (aaDist2) to 33 (LG model); For ANC4, different sites are from 32 (DNA model) to 47 (equally for WAG and LG models). To better understand the effects of the difference from computational inference, we extracted the sites that were not inferred by

any of the models we used by comparing with Hobbs' ANCs. The positions of the unique sites by my ASR after correcting gaps with '3U1H' are recorded as follows, for ANC1, the sites are 15, 19, 64, 110, 124, 149, 151, 153, 154, 156, 199, 322, 341, 355, and 358; For ANC2, the sites are 65, 149, 151, 176, 179, 319, 336, 355, and 359; For ANC3, the sites are 149, 151, 155, 179, 312, 319, 336, 337, 345, 355, 356, 358, and 359; And for ANC4, the sites are 84, 149, 151, 155, 255, 336, 341, 345, 350, 352, 353, 355, 356, 358, 359, 361, and 363. None of these sites in ANCs are in the functional sites (NAD binding, Magnesium/Manganese binding, substrate binding, and important catalysis sites [180]), except ANC4 at position 84 in NAD binding region. Overall, our ancestral inference differs substantially compared with Hobbs', and we could not repeat Hobbs' ancestral sequence reconstruction of LeuB computationally. However, the effects of these uniquely inferred sites on thermophily require further experimental investigations.

Conclusion

In summary, an uncertain phylogenetic tree, uncertain inferred ancestral sequences, and inconsistent reconstructed ancestral sequences confound analyses and predictions. All of these can lead to different phenotypic features for thermophily, and the explanation for the origin and evolution of thermophily would be totally different when using a different tree topology. Further, the effects of our reconstructed ancestral sequences on thermophily have to be investigated experimentally.

Abbreviation

ASR, Ancestral sequence reconstruction; LeuB, 3-isopropylmalate dehydrogenase.

Acknowledgements

This work was supported by grants to Dr. Eric A. Gaucher. We thank to the technical support from ‘The Partnership for an Advanced Computing Environment (PACE)’ at Georgia Institute of Technology.

CHAPTER 8

CONCLUSION

This dissertation presents my work that exploits phylogenetics and ancestral reconstruction to better understand genome evolution for both the functional divergence within individual gene families on the small-scale as well as gene content/organization at the genomic level on the large-scale. These small-scale studies focus on two gene families, thioredoxin and catenin, intended to deepen our understanding of both protein adaptation and innovation of new gene families through duplication events, respectively. Alternatively, the large-scale studies focus on both reassortments as revealed by diverse genotypes of H5N1 avian influenza viruses as well as inferences of gene content and genome rearrangements as revealed by ancestral genome content of a hypothetical ancient Mycoplasma species.

Genome evolution on both small-scale and large-scale was investigated for highly pathogenic H5N1 avian influenza viruses circulating in East Asia from 1996 to 2007 (Chapter Two and Three). Genotypes were assigned based on reassortment events within the gene pool of avian influenza viruses, with each genotype represented by eight precursor virus strains usually isolated before 1996. The precursors indicate the source of evolutionary origins for each of the eight separate gene segments comprising the whole genome of the virus. Altogether, 21 genotypes were identified for highly pathogenic H5N1 viruses circulating in East Asia from 1996 to 2007, all genotypes except two have two surface proteins HA and NA originate from a non-pathogenetic H5N1 strain

Tk/England/50-92/91 sequenced by us, and two genotypes include viruses affecting both avian and human hosts having the ability to cross the avian host barrier to human. Genome evolution on both large-scale reassortment and small-scale sequence mutations were found by their association with virus pathogenicity and host specificity. Particularly for H5N1 viruses emerging in 2001 and circulating from 2001 and 2007 in Vietnam, nine total genotypes were identified by frequent ressortment events of gene segments with precursor viruses isolated from China (Chapter Three). These various genotypes first emerged from north of Vietnam and then spread to south of Vietnam and further reassorted within local viral gene pools. In summary, diverse genotypes, along with viral isolation time and geographic locations, helped us understand viral emergence and patterns of transmission. Furthermore, the emergence, circulation, and transmission of viruses were associated with local ecological environments, geographic locations of isolation, and flight paths of migratory birds. These studies may be useful toward pandemic preparedness in virus surveillance and vaccine design, and in avian influenza prevention and control.

Functional divergence arising from genome evolution at the small-scale was studied in a cancer-related catenin gene family that functions in cell-cell adhesion and signaling pathways (Chapter Four). Our phylogenetic analysis showed that catenins arose during the origin of metazoans, underwent an array of duplication events along the metazoan history and experienced functional divergence leading to particular developmental physiologies and tissue-specific expression along with other co-evolved proteins. Furthermore, we resolved an annotation issue of a gene family based on biological functions of alpha catenin subfamily members, which are not homologs of beta and p120

subfamilies according to our analysis based on both sequence and structure similarity. We also resolved extensive gene annotation issues within the catenin gene family. These annotation resolutions highlight the application of evolutionary analysis in gene family annotation, and annotations of gene members within a gene family having either too close or too distant sequences and hard to resolve merely based on sequence similarity.

Small-scale genome evolution was further explored using a gene family called thioredoxin (a reductase enzyme conserved in three domains of life) by exploiting phylogenetics to reconstruct ancestral states to determine if paleoenvironments inferred from resurrected ancient thioredoxins are consistent with those found by the Gaucher group (Chapter Five). Ancestral states were inferred computationally based on a maximum likelihood algorithm with explicit evolutionary models for a series of ancestral nodes that interested us, including the last bacterial common ancestor, last eukaryote common ancestor, last archaea common ancestor and others. These ancestral nodes refer to a series of geological time points from the Precambrian era, and resurrected ancient genes were investigated biochemically for temperature and pH preference along geological time. The temperature trend achieved by ancient thioredoxin is consistent with that inferred using EF-Tu by the Gaucher group. In addition, ancient thioredoxin demonstrated high enzymatic activities under acidic conditions. Therefore, the temperature and pH inferred by computationally reconstructing and then chemically synthesizing ancient thioredoxins indicate a hot and acidic ancient environment that host ancient organisms in the Precambrian.

Large-scale genome evolution of both gene content and genome rearrangement was studied to both understand the evolution of Mycoplasma species as well as an attempt to

provide a reconstructed ancient genome to the synthetic biology community by taking advantage of technological breakthroughs in synthetic genomics and genome transplantation (Chapter Six). This study provided insights into genome evolution by linking genotypes and phenotypes to identity an innovation of lineage-specific pathogenicity. Our attempt of using ancestral genome reconstruction to generate a synthetic genome overcame a number of limitations associated with the conventional minimal genome method, and also provided a list of genes that were failed to be detected as essential by the minimal genome method due to limitations in the methodology such as the inability to incorporate duplicated genes, rich media and supplements causing non-essentiality of some metabolic genes, one-generation culturing due to instability of transposon mutagenesis, and low coverage transposon interruption density due to its dependence on gene length. We expect our computationally-inferred ancient gene content, along with technological breakthroughs of genome manipulation by the J. Craig Venter Institute, will benefit the synthetic biology community in creating an ancient or engineered organism for applications of biomedicine and biofuel.

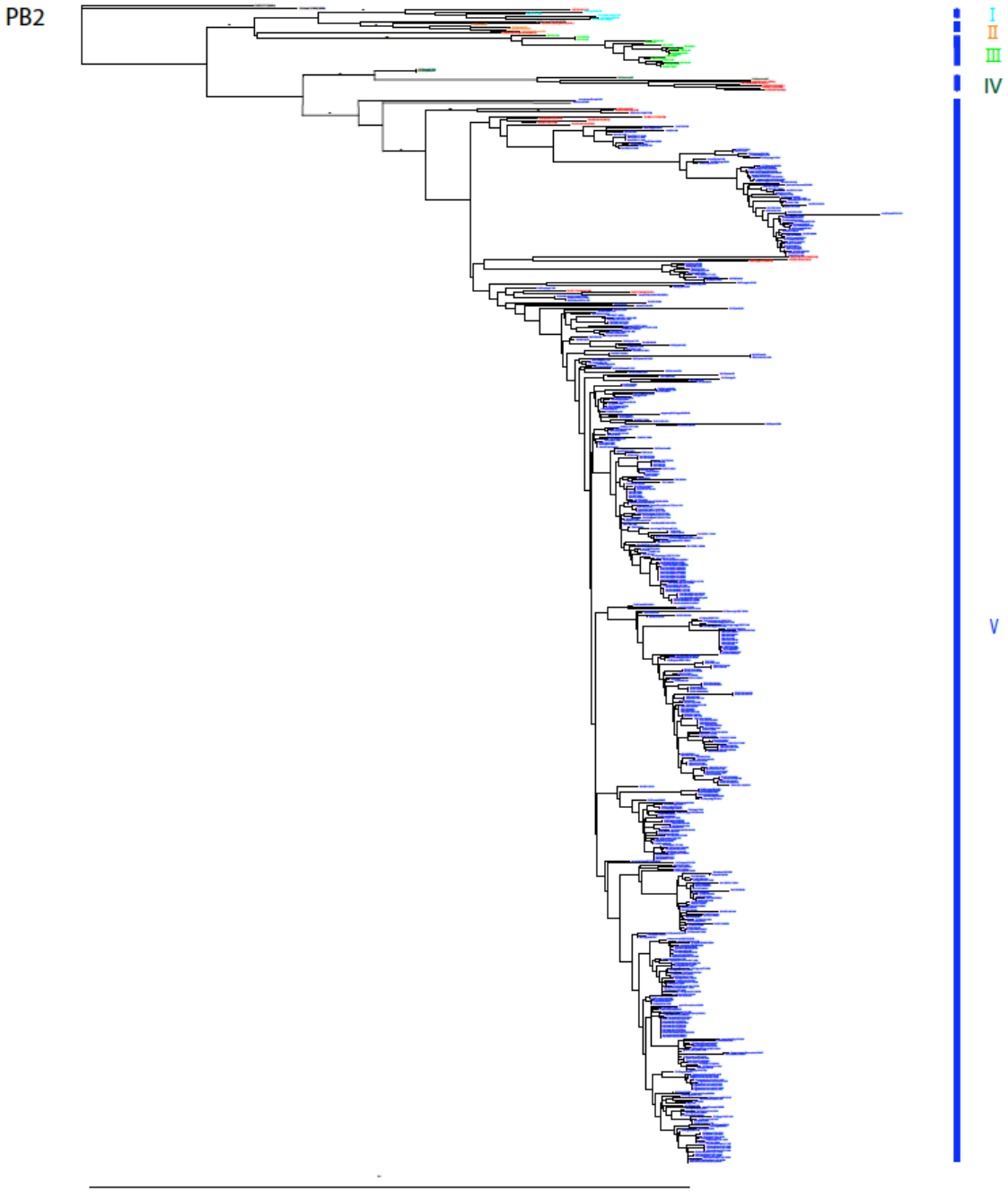
Lastly, computational analyses were performed to validate the performances of different ancestral sequence reconstruction methods by assessing accuracy and correctness of ASR using 3-isopropylmalate dehydrogenase (LeuB) of *Bacillus* (Chapter Seven). Our computational simulations indicate that the accuracy and correctness of ASR are dependent on computational algorithms (Maximum Parsimony and Maximum Likelihood integrated with various evolutionary models), types of input datasets (DNA, codon and amino acid), and different phylogenetic topologies. Our studies of LeuB phylogenetics, ASR and computational simulations failed to support Hobbs' resurrection

of LeuB of *Bacillus*, which indicates the presence of an alternative evolutionary pattern of thermophily by LeuB in *Bacillus*. However, experimental validations of our reconstructions and a confident LeuB phylogeny are required in follow-up studies. In total, this particular computational work provides us with greater insights to the accuracies and limitations of ancestral sequence reconstruction methods.

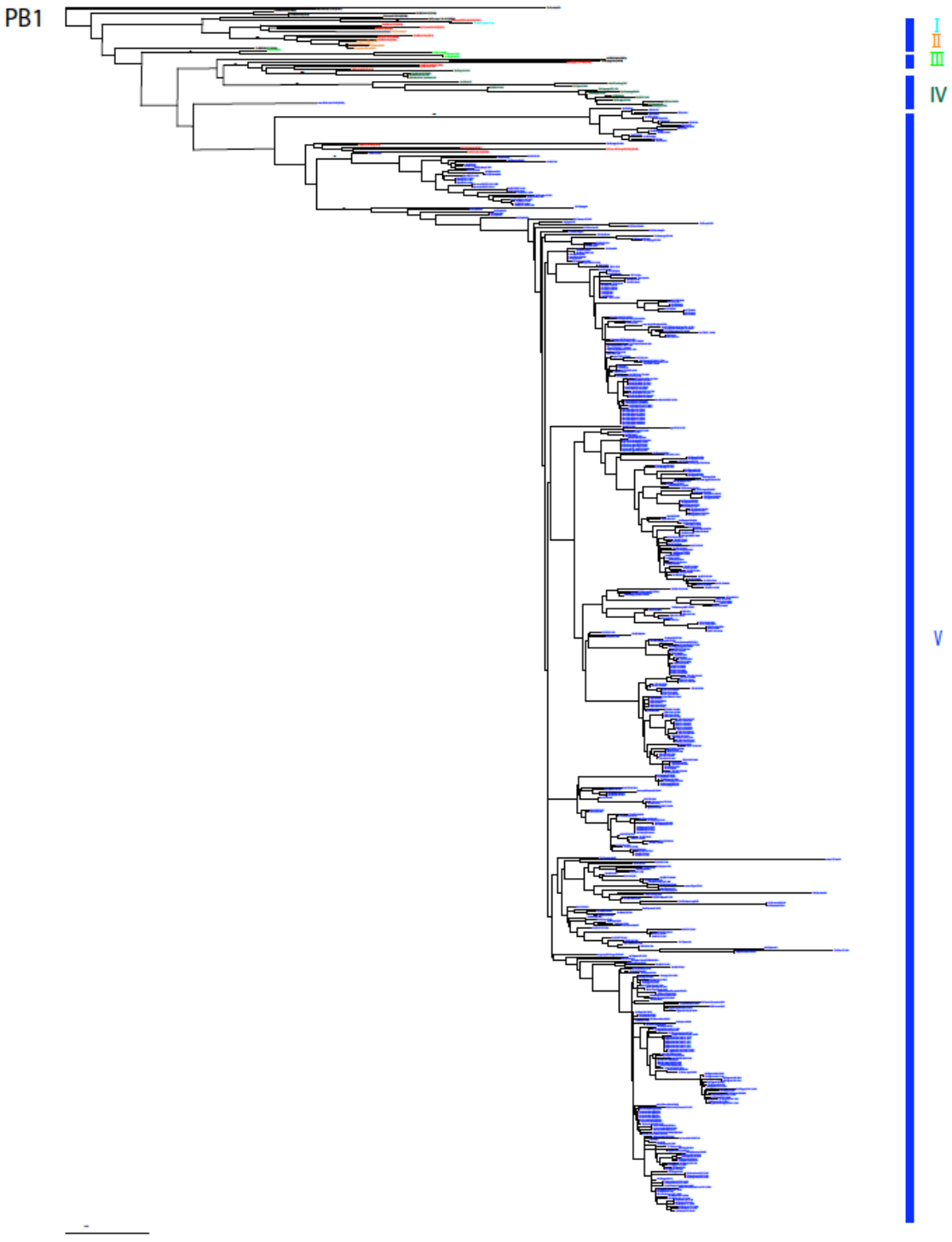
Overall, the work in this dissertation exploits phylogenetics and ancestral reconstruction to understand genome evolution at both the small-scale and the large-scale using modern and ancient organisms. Our studies highlight the diverse questions that evolutionary studies attempt to address and the different biological levels that can be studied to answer these questions.

APPENDIX A

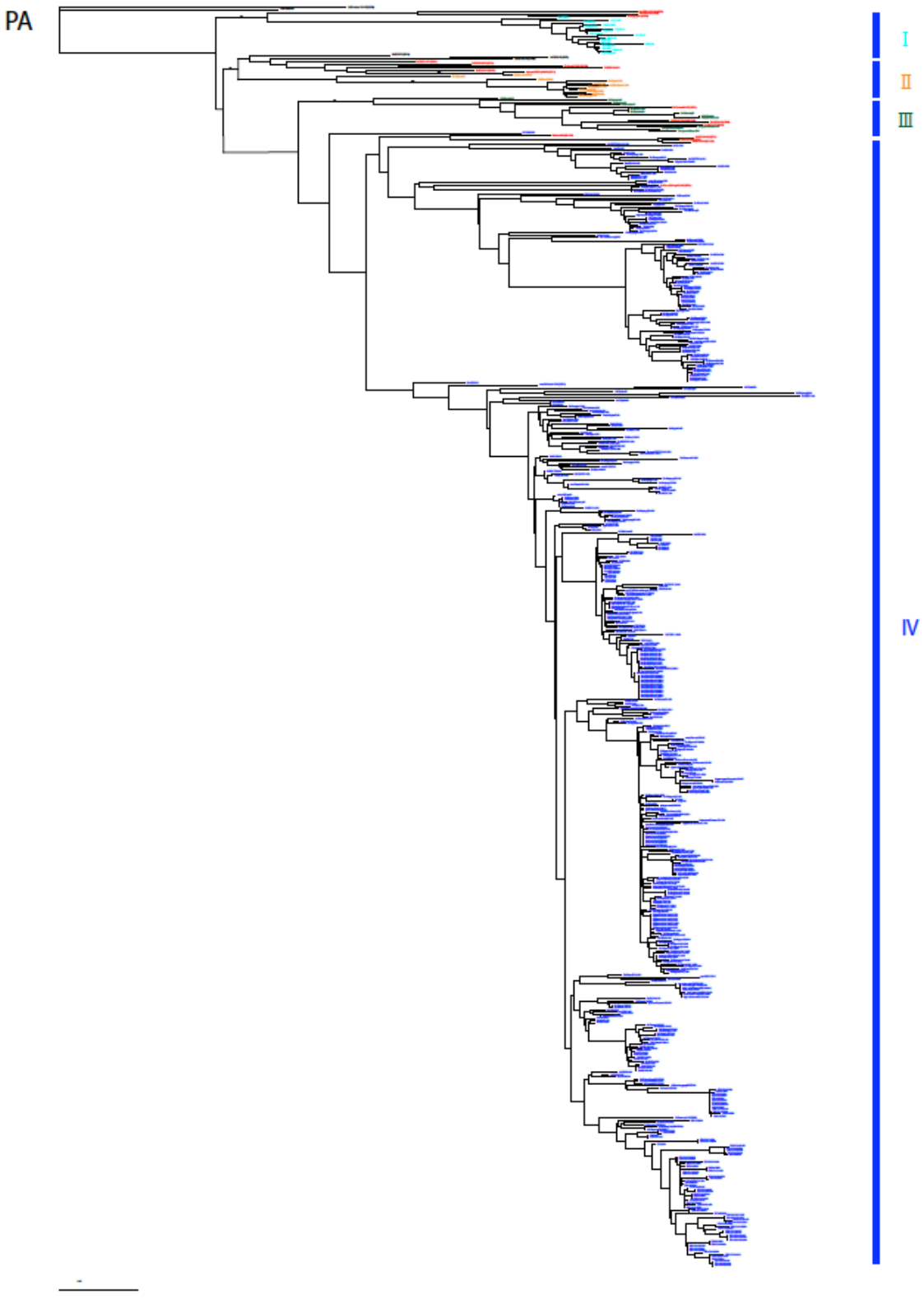
SUPPLEMENTARY INFORMATION FOR CHAPTER 2



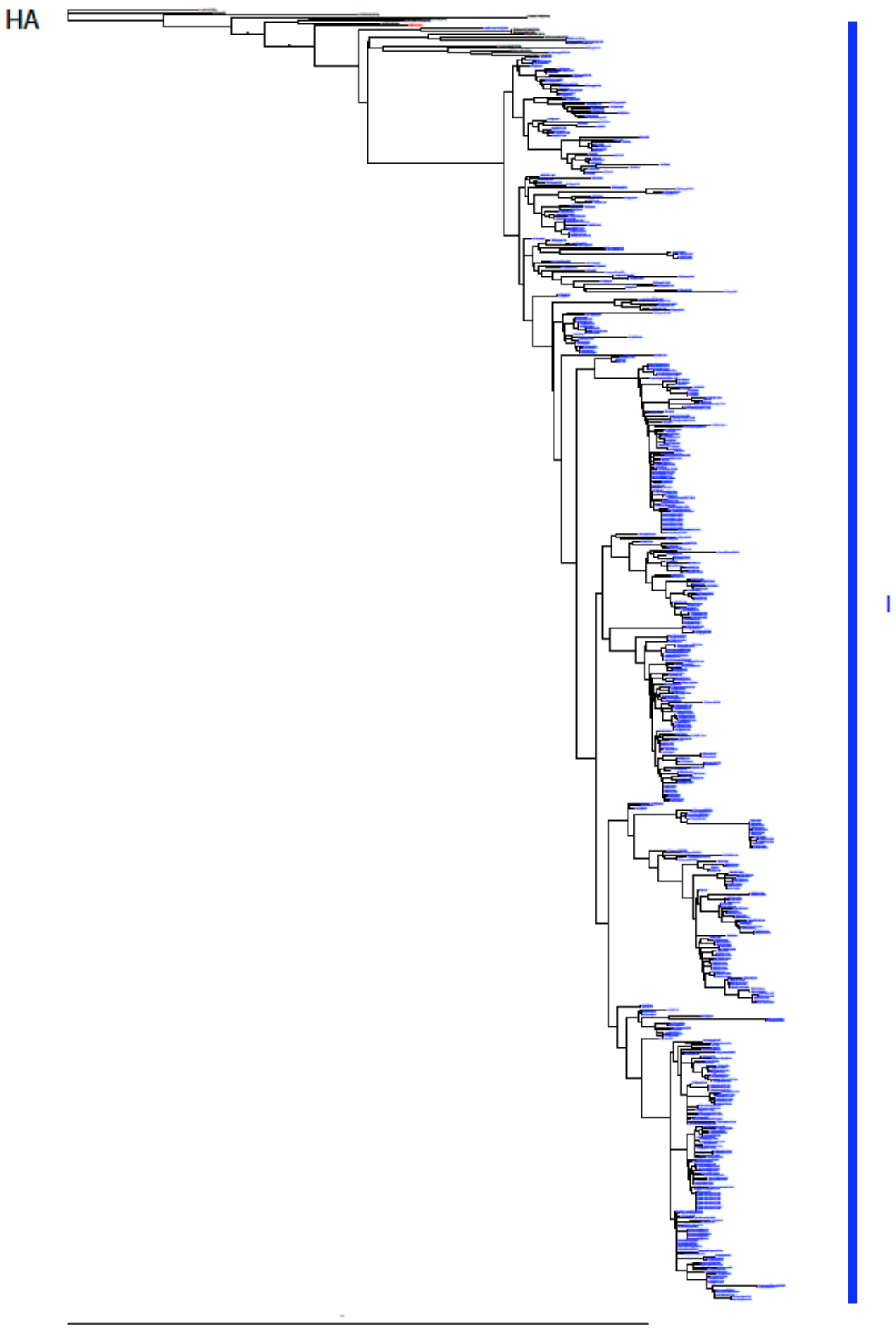
Supplementary figure A.1: Phylogenetic trees of PB2, PB1, PA, HA, NP, NA, MP, and NS for H5 HPAIVs and associated progenitor genes.



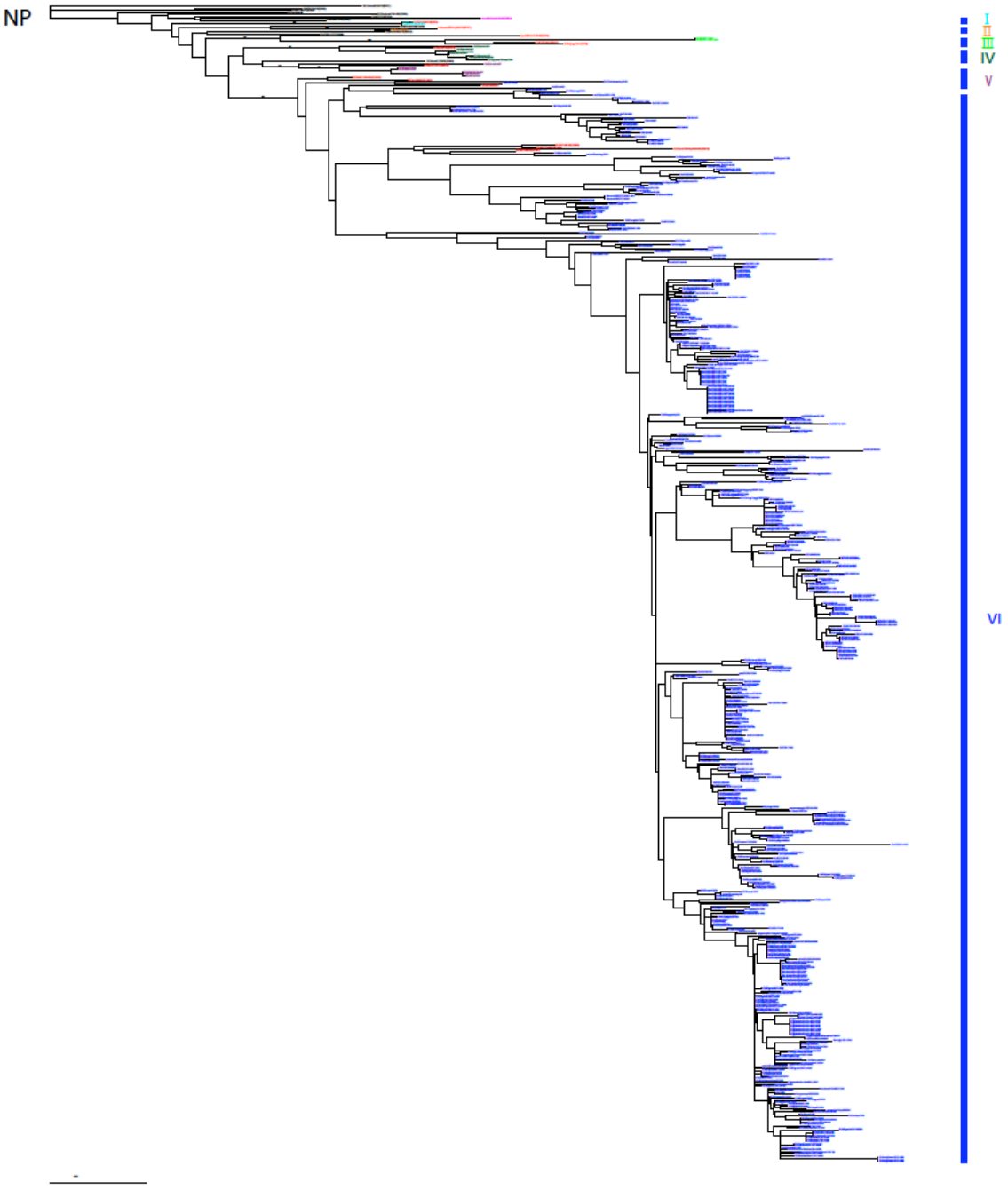
Supplementary figure A.1 continued



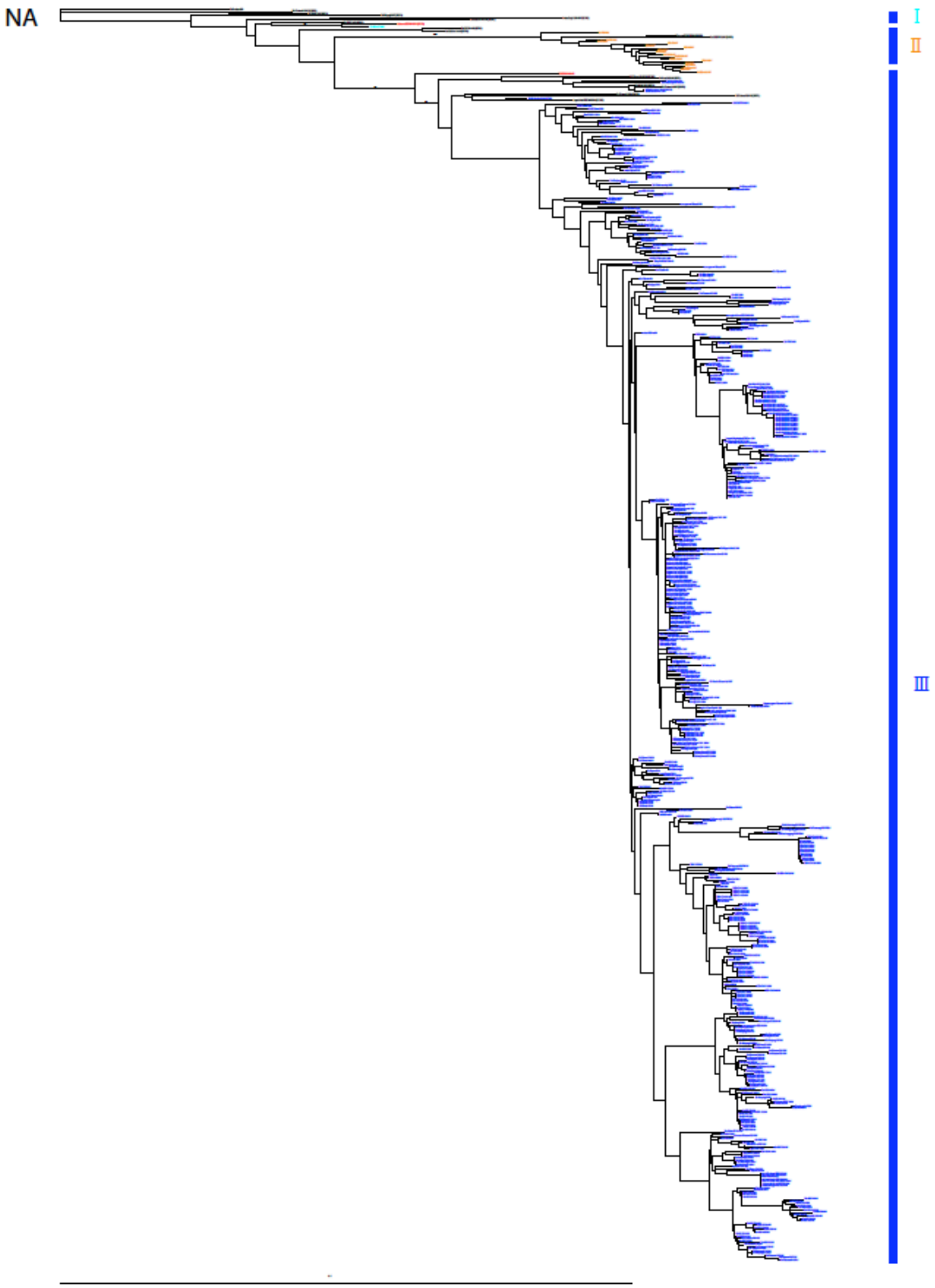
Supplementary figure A.1 continued



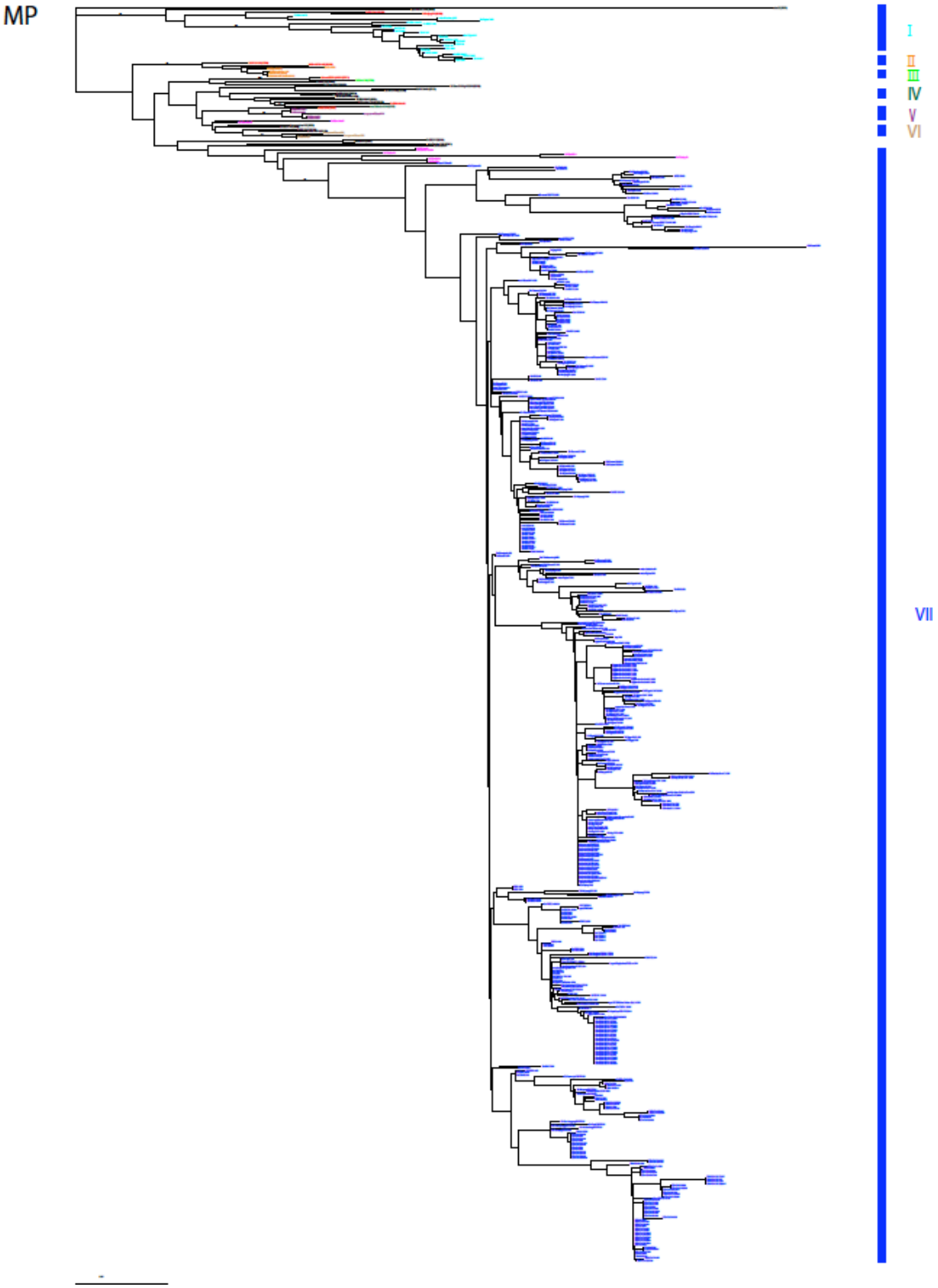
Supplementary figure A.1 continued



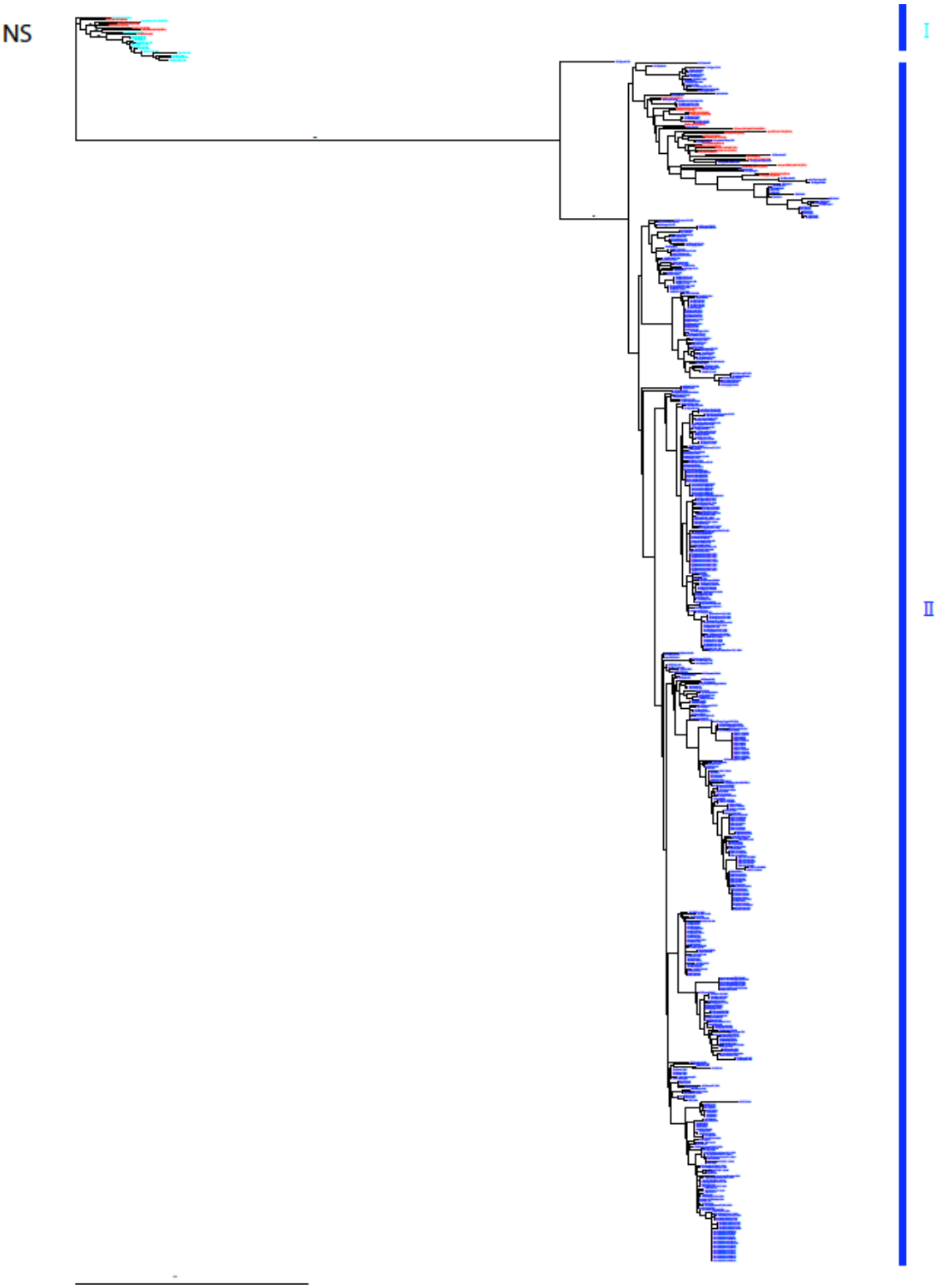
Supplementary figure A.1 continued



Supplementary figure A.1 continued



Supplementary figure A.1 continued



Supplementary figure A.1 continued

Supplementary table A.1: Reassortants generated from the putative precursor genes identified in this study.

*NA denotes the gene sequence is not available

Avian Influenza Virus	Gene segments*								Reassortant
	PB2	PB1	PA	HA	NP	NA	MP	NS	
Dk/GD/07/00	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Dk/GD/12/00	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Dk/SH/08/01	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Dk/ZJ/11/00	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Env/HK/437.10/99	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Env/HK/437.4/99	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Env/HK/437.6/99	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Env/HK/437.8/99	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Gs/GD/1/96	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Gs/GD/3/97	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Gs/HK/ww26/00	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Gs/HK/ww28/00	V	V	IV	I	VI	III	VII	I	H5N1-PR1
Ck/Hebei/718/01	V	V	IV	I	II	I	VII	II	H5N1-PR10
Ck/HK/FY150/01	V	V	IV	I	III	III	VII	II	H5N1-PR11
Ck/HK/FY150/01.MB	V	V	IV	I	III	III	VII	II	H5N1-PR11
Dk/Hokkaido/Vac.1/04	V	IV	IV	I	VI	III	II	II	H5N1-PR12
R/Dk/MG/54/01.Dk/MG/47/01	V	IV	IV	I	VI	III	II	II	H5N1-PR12
Dk/FJ/13/02	V	IV	IV	I	VI	III	VII	II	H5N1-PR13
Dk/SH/37/02	V	IV	IV	I	VI	III	VII	II	H5N1-PR13
Gs/GD/xb/01	V	IV	IV	I	VI	III	VII	II	H5N1-PR13
Ck/Jilin/xw/03	V	V	II	I	VI	III	VII	II	H5N1-PR14
Dk/GX/53/02	V	V	II	I	VI	III	VII	II	H5N1-PR14
Dk/SH/35/02	V	V	II	I	VI	III	VII	II	H5N1-PR14
WildDk/Hunan/211/05	V	V	II	I	VI	III	VII	II	H5N1-PR14
Ck/HK/31.2/02	I	IV	II	I	VI	III	VII	II	H5N1-PR15
Ck/Hubei/wf/02	V	V	IV	I	VI	III	I	II	H5N1-PR16
Ck/Hubei/wo/03	V	V	IV	I	VI	III	I	II	H5N1-PR16
Dk/GD/40/00	V	V	IV	I	VI	III	II	II	H5N1-PR17
Treesparrow/Henan/4/04	V	II	III	I	IV	III	V	II	H5N1-PR18
swine/GX/wz/04	V	V	IV	I	IV	III	VII	II	H5N1-PR19
Ck/HK/258/97	III	V	I	I	VI	II	I	II	H5N1-PR2
Ck/HK/728/97	III	V	I	I	VI	II	I	II	H5N1-PR2
Ck/HK/786/97	III	V	I	I	VI	II	I	II	H5N1-PR2
Dk/HK/p46/97	III	V	I	I	VI	II	I	II	H5N1-PR2
Dk/VN/1/05	III	V	I	I	VI	II	I	II	H5N1-PR2
Dk/VN/8/05	III	V	I	I	VI	II	I	II	H5N1-PR2
Gs/HK/w355/97	III	V	I	I	VI	II	I	II	H5N1-PR2
Gs/VN/3/05	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/156/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/481/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/483/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/485/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/486/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/532/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/538/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/542/97	III	V	I	I	VI	II	I	II	H5N1-PR2
HK/97/98	III	V	I	I	VI	II	I	II	H5N1-PR2
Md/Italy/3401/05	I	IV	IV	I	VI	III	II	II	H5N1-PR20
Treesparrow/Henan/2/04	II	II	III	I	IV	III	VI	II	H5N1-PR20
Treesparrow/Henan/3/04	II	II	III	I	IV	III	VI	II	H5N1-PR20
Ck/Hubei/wl/97	II	III	III	I	IV	III	V	II	H5N1-PR3
Ck/Henan/210/04	IV	III	III	I	V	III	V	II	H5N1-PR4
Ck/Henan/wu/04	IV	III	III	I	V	III	V	II	H5N1-PR4
Ck/Hubei/wj/97	IV	III	III	I	V	III	V	II	H5N1-PR4
Ck/Hubei/wh/97	V	II	III	I	V	III	VI	II	H5N1-PR5
Dk/Yokohama/aq10/03	I	V	IV	I	VI	III	VII	II	H5N1-PR6

Supplementary table A.1 continued

Gs/ST/2086/06	I	V	IV	I	VI	III	VII	II	H5N1-PR6
Gs/ST/239/06	I	V	IV	I	VI	III	VII	II	H5N1-PR6
Bar.headedGs/MG/1/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/0510/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/12/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/5/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/59/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/60/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/61/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/62/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/65/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/67/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/68/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Bar.headedGs/QH/75/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Beijing/01/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
bird/TH/3.1/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Brown.headedGull/QH/3/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
buzzard/DM/6370/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
cat/Germany/606/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
cat/TH/KU.02/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
China/GD01/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
China/GD02/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Afg/1207/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Afg/1573.47/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Afg/1573.65/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Afg/1573.7/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Afg/1573.92/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Ayutthaya/TH/CU.23/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/BurkinaFaso/13.1/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/BurkinaFaso/1347.16/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/CotedIvoire/1787.34/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Crimea/08/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Dairi/BPPVI/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/DeliSerdang/BPPVI/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Egypt/2253.1/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Egypt/3/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/FJ/10039/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/FJ/1042/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/FJ/11933/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/FJ/12239/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/FJ/584/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/FJ/9821/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GD/174/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GD/178/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GD/191/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/2147/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/2173/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/29/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/3055/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/3570/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/3721/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/4059/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Guiyang/441/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GunungKidal/BBVW/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GX/12/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GX/1951/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GX/3154/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GX/3791/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GX/463/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/GX/4989/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

Ck/GX/604/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Hebei/326/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Henan/01/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Henan/12/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Henan/13/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Henan/16/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/282/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/31.4/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/715.5/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/947/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/96.1/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/NT873.3/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/NT873.3/01.MB	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/YU22/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/YU562/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/YU777/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/YU822.2/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/HK/YU822.2/01.MB	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Hubei/327/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Hubei/489/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Hunan/999/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/IND/BL/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/IND/CDC25/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/IND/PA/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/IvoryCoast/1787.35/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/JianGsu/cz1/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Jilin/hh/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Jilin/xv/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Korea/ES/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Krasnodar/123/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Krasnodar/300/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Kurgan/05/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Kurgan/3/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Kyoto/3/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Magetan/BBVW/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Moscow/2/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nakorn.Patom/TH/CU.K2/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/1047.30/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/1047.34/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/1047.54/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/1047.62/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/1047.8/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/641/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/957.20/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/AB13/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/AB14/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/BA210/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/BA211/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/FA7/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/IF10/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/OD9/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/SO300/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/SO452/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/SO493/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Nigeria/SO494/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Oita/8/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Omsk/14/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Purworejo/BBVW/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Russia/Krasnodar/2/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/ST/1233/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

Ck/ST/3840/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/ST/3923/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Shanxi/2/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Simalanggang/BPPVI/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/ST/4231/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Sudan/1784.10/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Sudan/1784.7/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Sudan/1784.8/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Sudan/2115.10/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Sudan/2115.12/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Sudan/2115.9/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Suzdalka/06/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Tarutung/BPPVI/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TebingTinggi/BPPVI/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TH/9.1/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TH/ICRC.V143/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TH/Kanchanaburi/Ck.160/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TH/Nontaburi/Ck.162/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TH/NP.172/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TH/PC.168/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/TH/PC.170/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Tula/4/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/10/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/11/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/17/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/2/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/33/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/35/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/36/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/37/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/38/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/6/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/8/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/9/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/C57/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/C58/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/LD.080/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/TG.023/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/VN/TN.025/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ck/Yamaguchi/7/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
commongoldeneye/MG/12/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
commonmagpie/HK/2125/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
commonmagpie/HK/2256/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
commonmagpie/HK/3033/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
commonmagpie/HK/645/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
crestedeagle/Belgium/01/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
crestedmyna/HK/540/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
crow/Kyoto/53/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
crow/Osaka/102/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnuscygnus/Iran/754/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnuscygnus/Krasnodar/329/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.1/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.10/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.2/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.3/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.4/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.5/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.6/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.7/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Ast/Ast05.2.8/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

Cygnusolor/Ast/Ast05.2.9/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Croatia/1/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Italy/742/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Cygnusolor/Italy/808/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/China/E319.2/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/CotedIvoire/1787.18/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Egypt/2253.3/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/10934/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/11094/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/11311/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/12032/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/17/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/668/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/671/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/897/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/9651/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/FJ/9713/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GD/173/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GD/22/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Guangzhou/20/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Guiyang/2231/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Guiyang/293/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Guiyang/3242/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Guiyang/3834/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Guiyang/3996/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Guiyang/497/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/07/99	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/13/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/1311/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/150/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/1793/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/1830/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/2143/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/22/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/2775/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/288/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/3085/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/3364/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/35/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/3548/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/3741/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/3819/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/392/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/4016/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/4184/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/4196/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/4428/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/4665/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/4830/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/50/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/5165/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/5270/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/5457/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/744/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/793/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/804/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/89/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/951/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/GX/xa/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/HK/2986.1/00	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

Dk/Hubei/wp/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hubei/wq/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/1265/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/1608/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/1652/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/324/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/344/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/5106/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/5152/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/856/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Hunan/988/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/IND/MS/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Korea/ESD1/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Kurgan/08/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Niger/914/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Novosibirsk/02/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Novosibirsk/56/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Parepare/BBVM/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/SH/13/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/SH/xj/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/ST/13323/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/ST/4610/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/VN/12/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/VN/18/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/VN/19/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/VN/20/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Yunnan/4589/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Yunnan/5236/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Yunnan/5251/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/Yunnan/5877/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/ZJ/52/00	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Dk/ZJ/bj/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
egret/HK/757.2/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
feline/IND/CDC1/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
GreatBlack.headedGull/QH/2/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Grebe/Novosibirsk/29/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Grebe/Tyva/Tyv06.1/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Grebe/Tyva/Tyv06.2/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Grebe/Tyva/Tyv06.8/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
greylagGs/DM/6692/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/FJ/bb/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Guiyang/337/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Guiyang/3422/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/1198/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/1458/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/1633/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/1898/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/2112/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/224/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/3017/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/3316/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/345/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/4289/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/4513/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/532/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/5414/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/GX/582/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/HK/3014.8/00	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Hungary/3413/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Jilin/hb/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

Gs/Krasnoozerskoe/627/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/ST/2216/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/ST/3265/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/ST/3295/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Suzdalka/10/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Yunnan/3315/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Yunnan/4129/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Yunnan/4494/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Yunnan/4804/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Gs/Yunnan/6027/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
guineafowl/Nigeria/957.12/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Guineafowl/ST/1341/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
HK/212/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
HK/213/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/160H/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/175H/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/239H/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/245H/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/283H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/292H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/298H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/304H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/341H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/5/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/534H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/535H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/538H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/546H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/560H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/567H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/569H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/583H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/604H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/7/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1031/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1031RE2/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1031T/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1031T2/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1032/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1032N/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1032T/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1046/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC1047/07	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC184/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC287E/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC292T/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC326/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC326N/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC326T/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC329/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC357/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC370/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC370E/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC390/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC523/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC523E/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC523T/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC582/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC594/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC595/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

IND/CDC596/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC597/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC599/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC599N/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC610/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC623/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC623E/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC624/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC624E/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC625/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC625L/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC634/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC634P/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC634T/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC644/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC644T/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC669/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC669P/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC699/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC7/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC739/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC742/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC759/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC835/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC836/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC836T/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC887/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC938/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC938E/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
IND/CDC940/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Iraq/1/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Japanesewhite.eye/HK/1038/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
large.billedcrow/HK/2512/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
leopard/Suphanburi/TH/Leo.1/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
littleegret/HK/718/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
littleGrebe/TH/Phichit.01/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Md/Italy/835/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
migratoryDk/Jiangxi/2300/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
munia/HK/2454/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Bangkok/LBD0111F/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Bangkok/LBD0511F/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBA0627May/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBA2611M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBA2811M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBA2911M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBA3011M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD0104F/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD0404F/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD1221J/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD1421J/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD1521J/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD1621J/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD1721J/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD1821J/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD2316F/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD2416F/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD2616F/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD3009M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD3209M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD3309M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

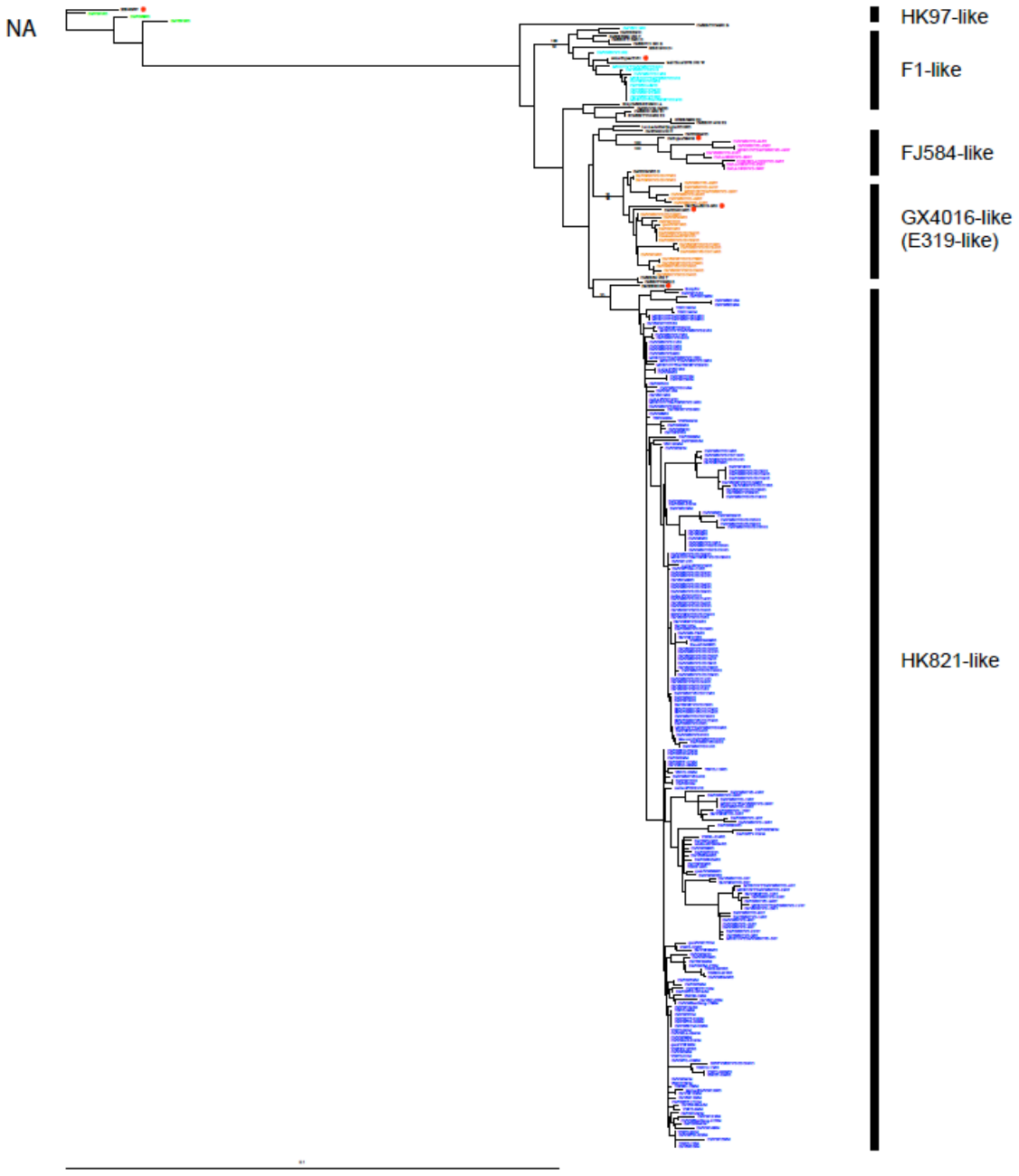
Obs/Nakh/BBD3509M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD3516M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD3616M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD4011M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD4211M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Nakh/BBD4411M/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Obs/Suphanburi/TSD0912F/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
ostrich/Nigeria/1047.25/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Owstonscivet/VN/1/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
peacoCk/DM/60295/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
peregrine/DM/6632/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
peregrinefalcon/HK/D0028/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ph/HK/FY155/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Ph/HK/FY155/01.MB	V	V	IV	I	VI	III	VII	II	H5N1-PR7
pheasant/ST/2239/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Qa/TH/57/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Qa/GX/575/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Qa/TH/NakhonPathom/QA.161/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Qa/VN/15/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Qa/VN/36/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Shenzhen/406H/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
SilkyCk/HK/SF189/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Swan/Germany/R65/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Swan/GX/307/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Swan/Slovenia/760/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
swine/Anhui/ca/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
swine/FJ/1/03	V	V	IV	I	VI	III	VII	II	H5N1-PR7
swine/FJ/F1/01	V	V	IV	I	VI	III	VII	II	H5N1-PR7
swine/Henan/wy/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Teal/China/2978.1/02	V	V	IV	I	VI	III	VII	II	H5N1-PR7
TH/1/KAN.1	V	V	IV	I	VI	III	VII	II	H5N1-PR7
TH/16/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
TH/2/SP.33	V	V	IV	I	VI	III	VII	II	H5N1-PR7
TH/5/KK.494	V	V	IV	I	VI	III	VII	II	H5N1-PR7
TH/676/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
TH/NK165/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
TH/SP83/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tiger/Suphanburi/TH/Ti.1/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tiger/TH/CU.T3/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tiger/TH/VSMU.1.SPB/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tk/12/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tk/Egypt/2253.2/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tk/IvoryCoast/4372.2/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tk/IvoryCoast/4372.3/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tk/IvoryCoast/4372.4/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tk/Suzdalka/12/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
Tk/Turkey/1/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
tuftedDk/DM/6431/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
tuftedDk/DM/6540/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
VN/1194/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
VN/1203/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
VN/3062/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
VN/CL26/04	V	V	IV	I	VI	III	VII	II	H5N1-PR7
white.baCkedmunia/HK/2469/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
whooperSwan/DM/7224/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
whooperSwan/DM/7275/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
whooperSwan/MG/2/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7
whooperSwan/MG/3/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
WildDk/Hunan/021/05	V	V	IV	I	VI	III	VII	II	H5N1-PR7
ZJ/16/06	V	V	IV	I	VI	III	VII	II	H5N1-PR7

Supplementary table A.1 continued

Ck/HK/37.4/02	V	IV	IV	I	VI	III	I	II	H5N1-PR8
Dk/FJ/19/00	V	IV	IV	I	VI	III	I	II	H5N1-PR8
swine/Shandong/2/03	V	IV	IV	I	VI	III	I	II	H5N1-PR8
Ck/Hebei/108/02	V	IV	II	I	VI	III	VII	II	H5N1-PR9
Dk/Anyang/AVL.1/01	V	IV	II	I	VI	III	VII	II	H5N1-PR9
Dk/FJ/01/02	V	IV	II	I	VI	III	VII	II	H5N1-PR9
Dk/Shandong/093/04	V	IV	II	I	VI	III	VII	II	H5N1-PR9
Dk/SH/38/01	V	IV	II	I	VI	III	VII	II	H5N1-PR9
Gf/HK/38/02	V	IV	II	I	VI	III	VII	II	H5N1-PR9
Swan/Hokkaido/51/96(H5N3)	V	V	IV	I	Unknown	NA	IV	I	O-PR1
Dk/Hokkaido/55/96(H1N1)	NA	NA	NA	NA	NA	III	NA	NA	O-PR2
Ck/Hebei/1/02(H7N2)	V	I	II	NA	I	NA	III	I	O-PR4
Dk/MG/54/01(H5N2)	V	IV	IV	I	VI	NA	II	I	O-PR3
Ck/Jilin/9/04	V	V	IV	I	VI	III	Unknown	II	Undefined
Ck/Jilin/ha/03	V	V	IV	I	VI	III	Unknown	II	Undefined
Ck/Jilin/hd/02	V	V	IV	I	VI	III	Unknown	II	Undefined
Ck/Jilin/hf/02	V	V	IV	I	VI	III	Unknown	II	Undefined
Ck/Jilin/hg/02	V	V	IV	I	VI	III	Unknown	II	Undefined
Ck/Hubei/wk/97	IV	II	III	I	IV	III	Unknown	II	Undefined
WildDk/GD/314/04	V	V	III	I	V	III	Unknown	II	Undefined
Md/GX/wt/04	V	V	III	I	V	III	Unknown	II	Undefined
Ck/Hubei/wi/97	II	NA	III	I	IV	NA	Unknown	II	Undefined

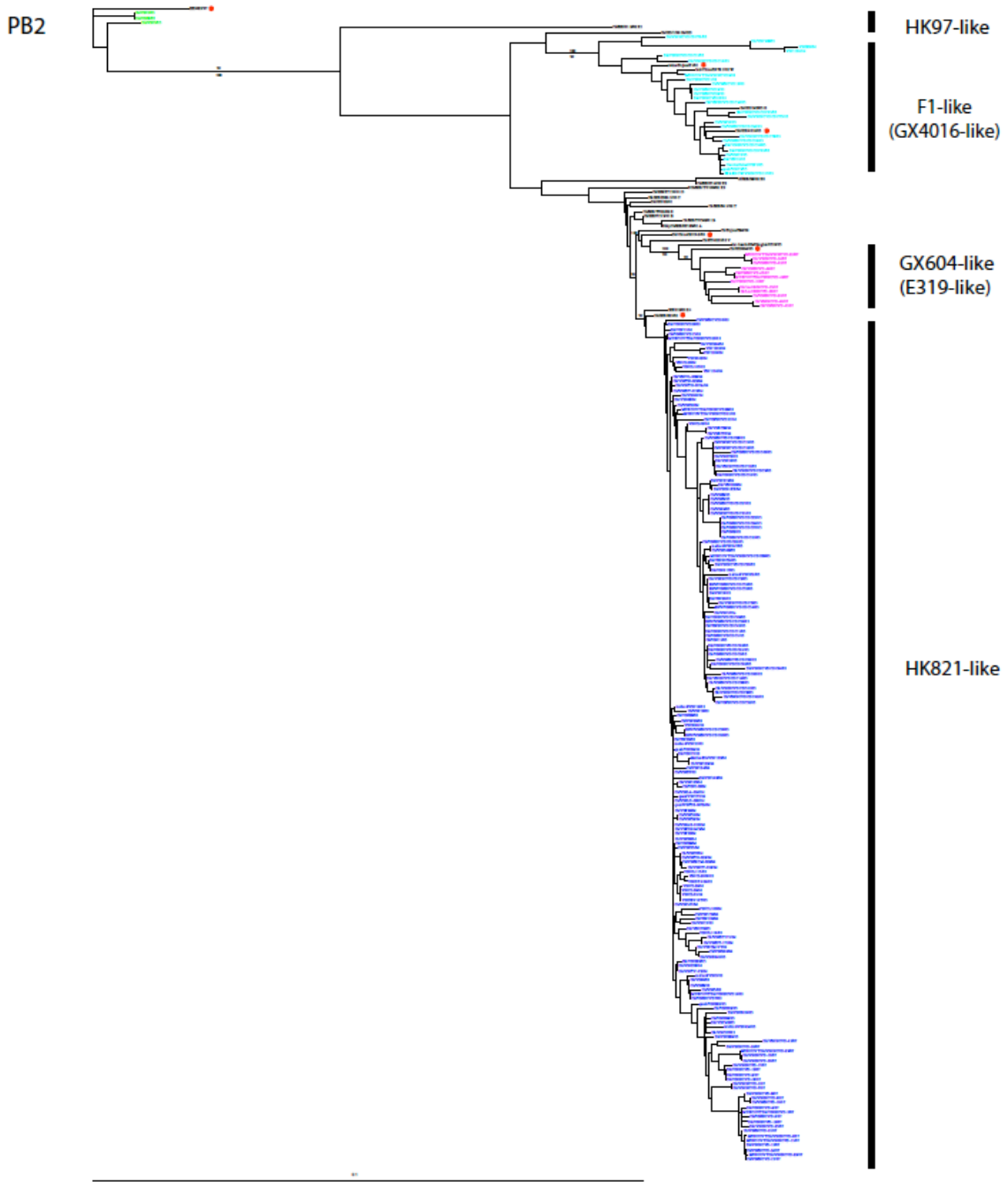
APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

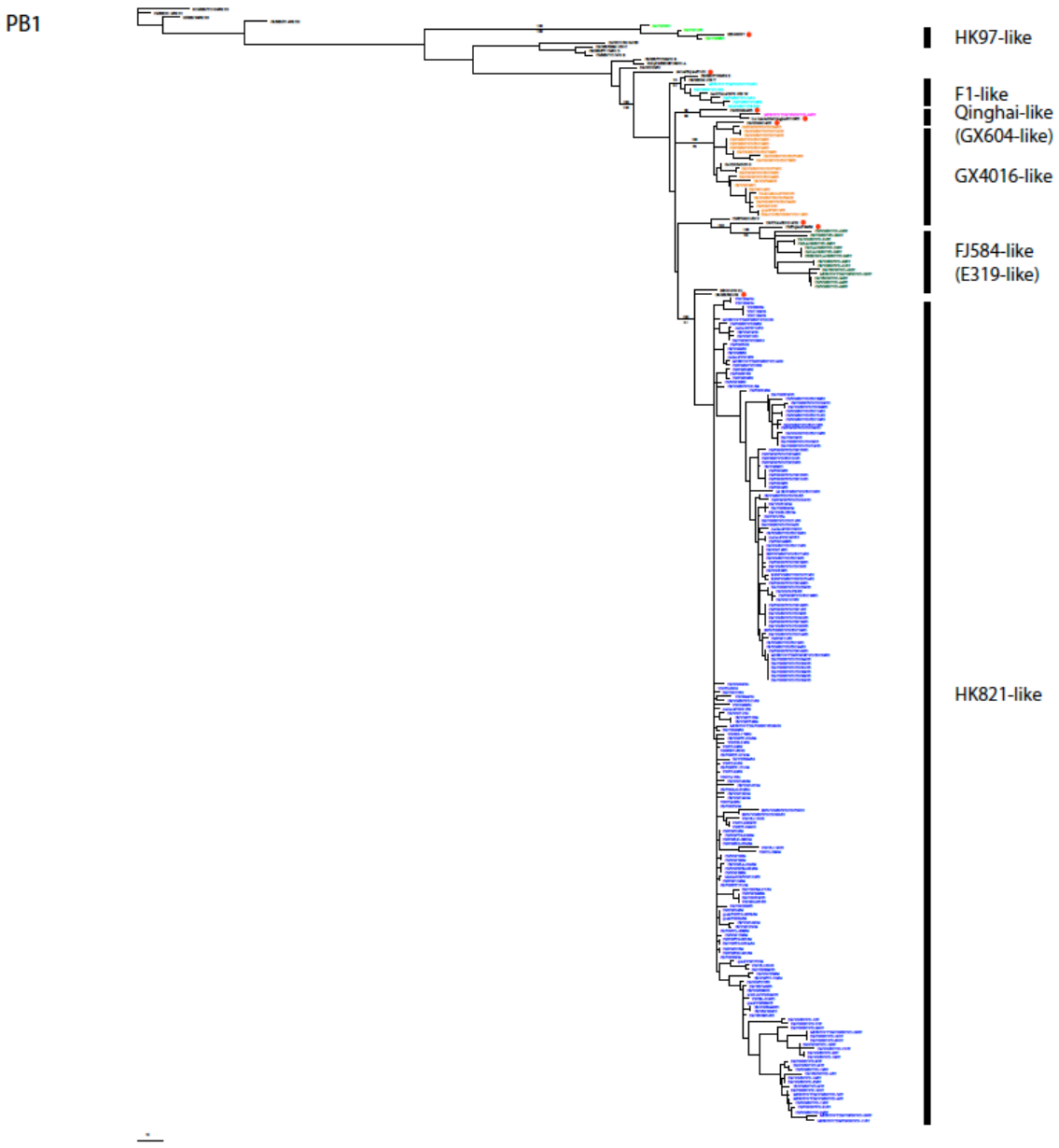


Supplementary figure B.1: Phylogenetic analyses of the H5N1 highly pathogenic avian influenza viruses (HPAIVs) isolated in Vietnam between 2001 and 2007.

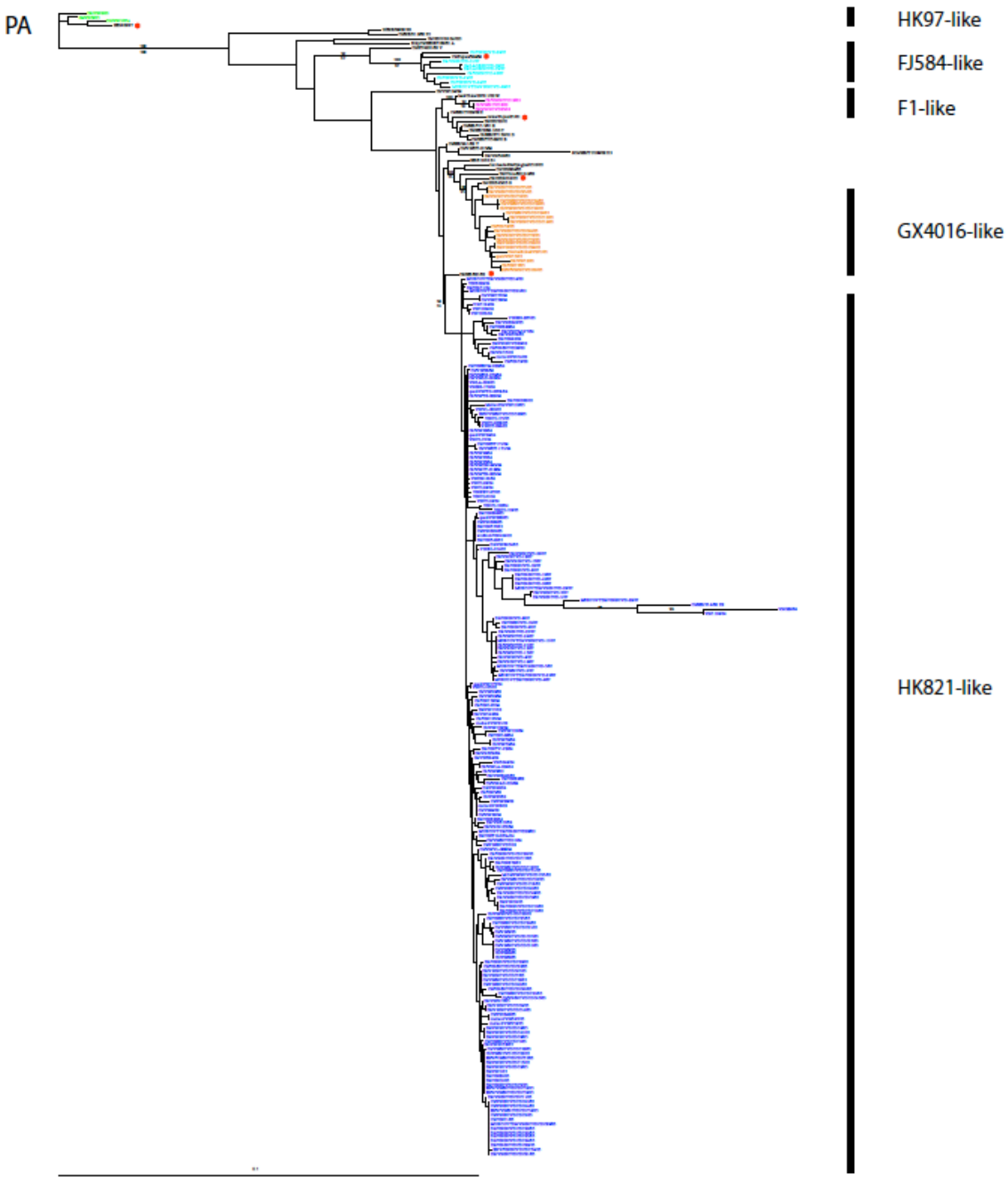
The figures are for segments NA, PB2, PB1, PA, NP, MP, and NS, respectively. Posterior probabilities and bootstrap values are given above and below branches, respectively. A red spot was marked besides each predicted precursor virus.



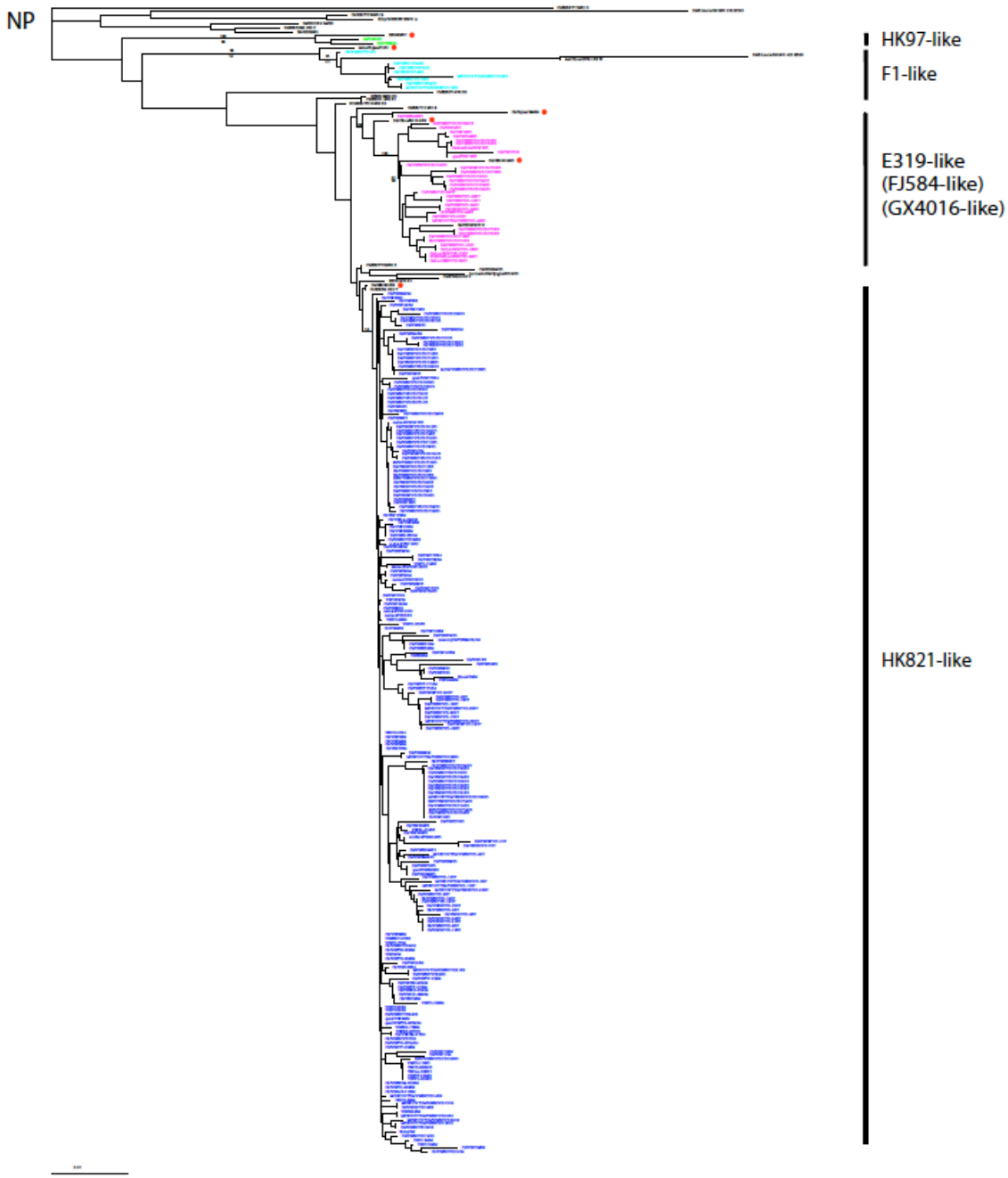
Supplementary figure B.1 continued



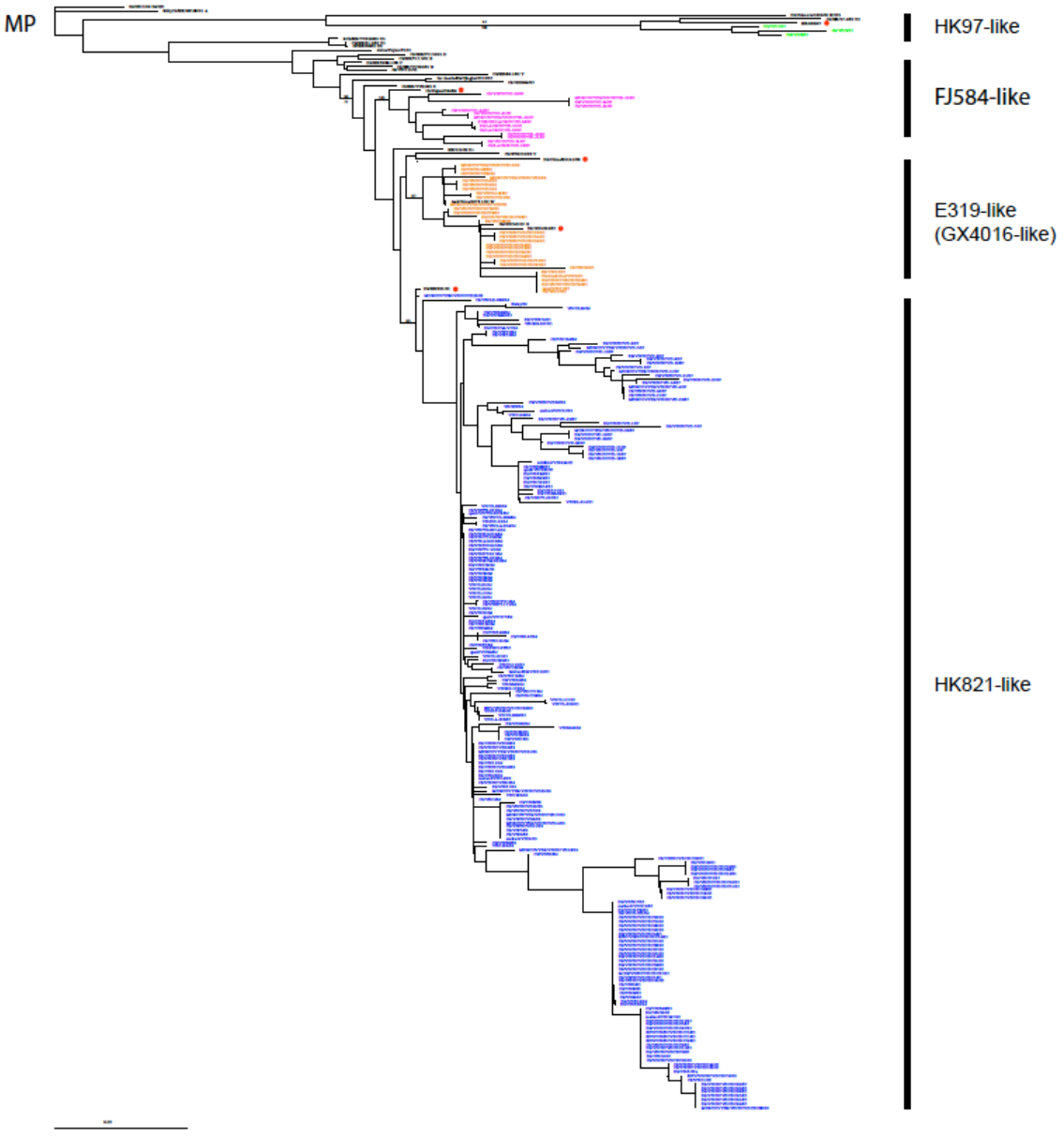
Supplementary figure B.1 continued



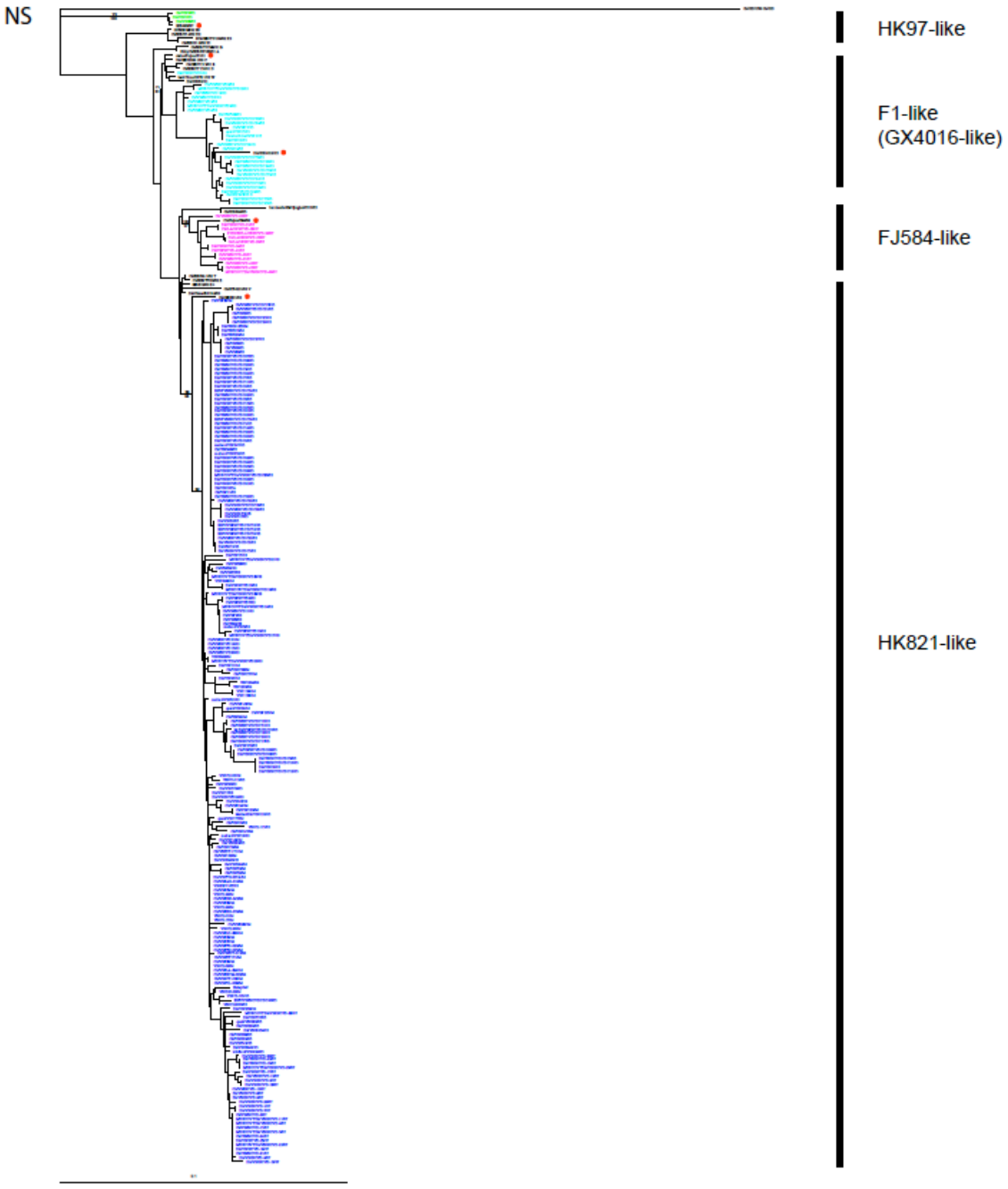
Supplementary figure B.1 continued



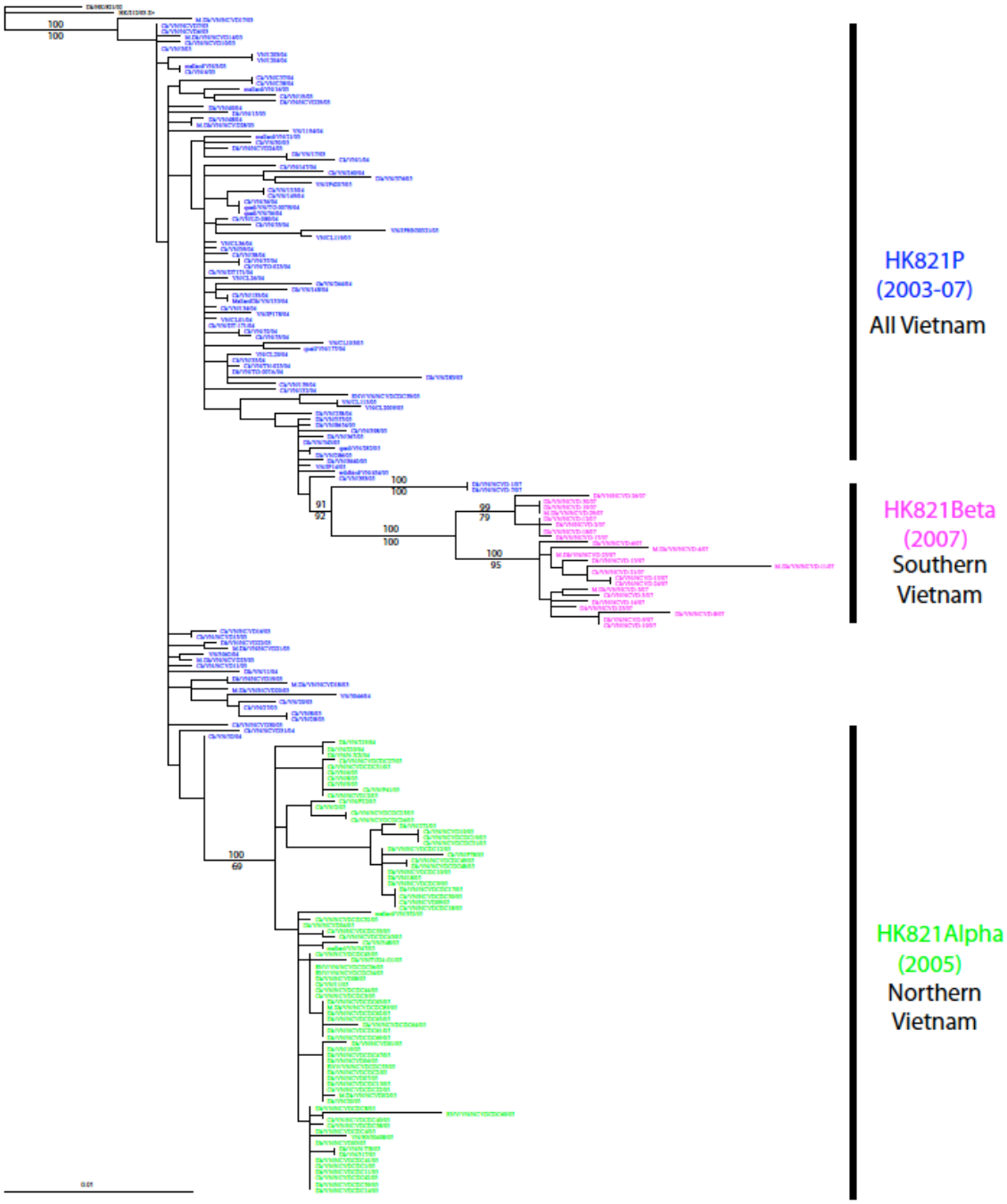
Supplementary figure B.1 continued



Supplementary figure B.1 continued



Supplementary figure B.1 continued

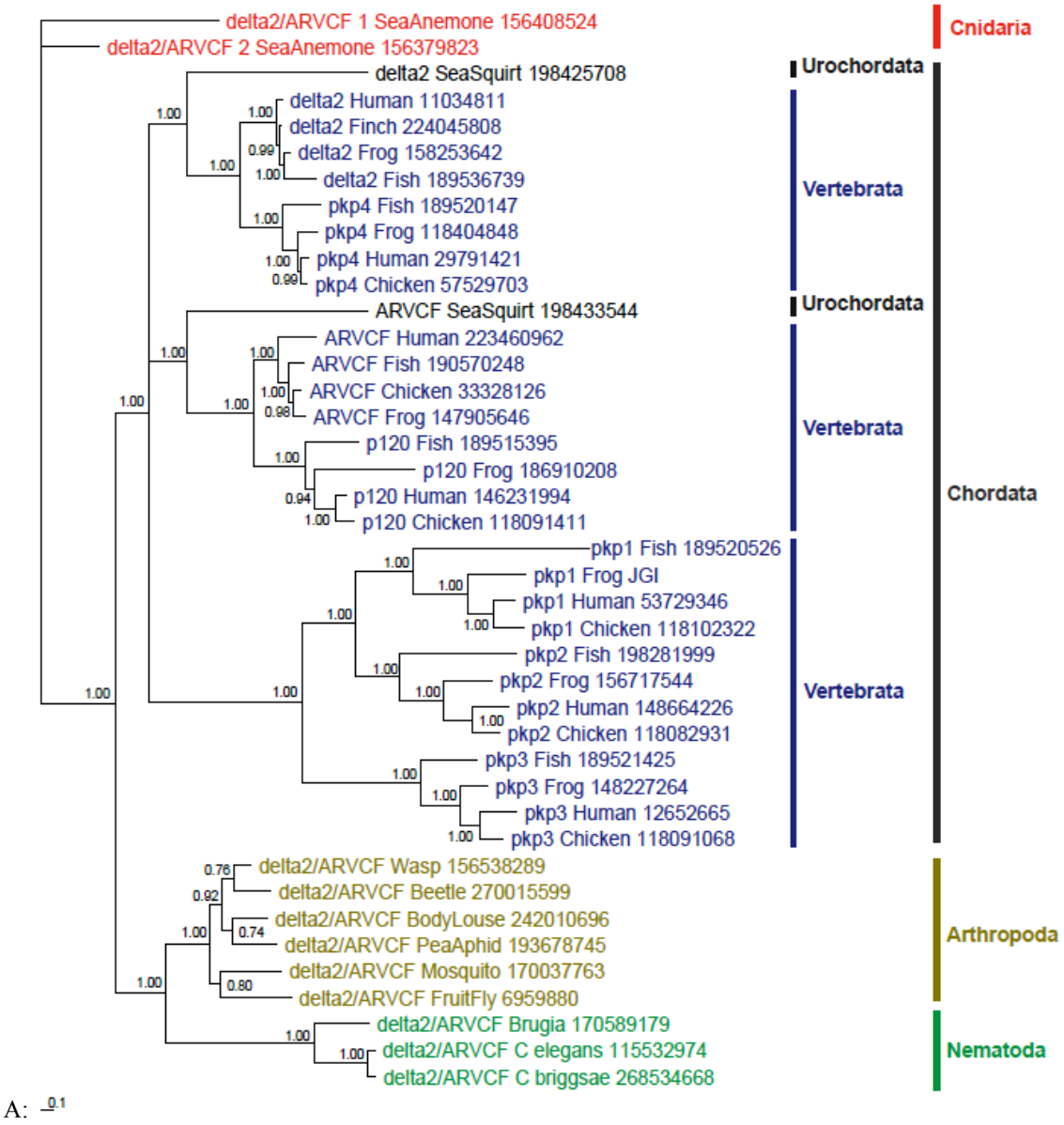


Supplementary figure B.2: HK821-like viruses formed three sub-lineages: HK821P, HK821α, and HK821β.

Phylogenetic tree for HA gene of HK821-like AIVs isolated from Vietnam, and the tree was rooted by Dk/HK/821/02 (H5N1). The phylogenetic inference was described in the legend of Figure 3.1.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR CHAPTER 4



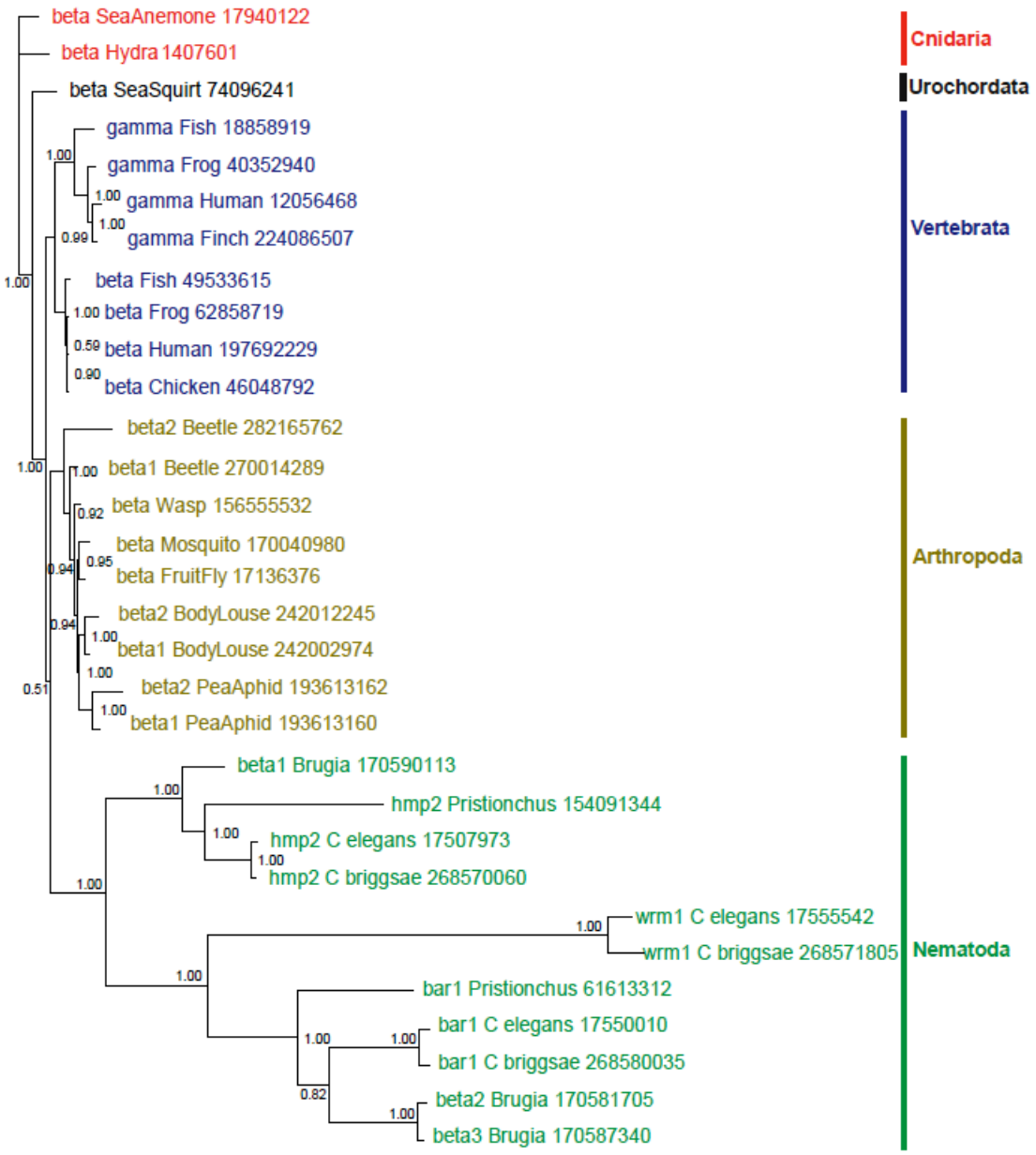
Supplementary figure C.1: Bayesian phylogenies for three subfamilies of the catenin family.

Bayesian phylogenies are shown with branch lengths and posterior probabilities, with each tree having sea anemone as the outgroup. Sequence/taxon names in the trees are a combination of gene name, species name, and NCBI sequence identifier. Similar color patterns were applied for species in different phyla as in Figure 4.1. For all phylogenies, scale bars represent amino acid replacements per site per unit evolutionary time.

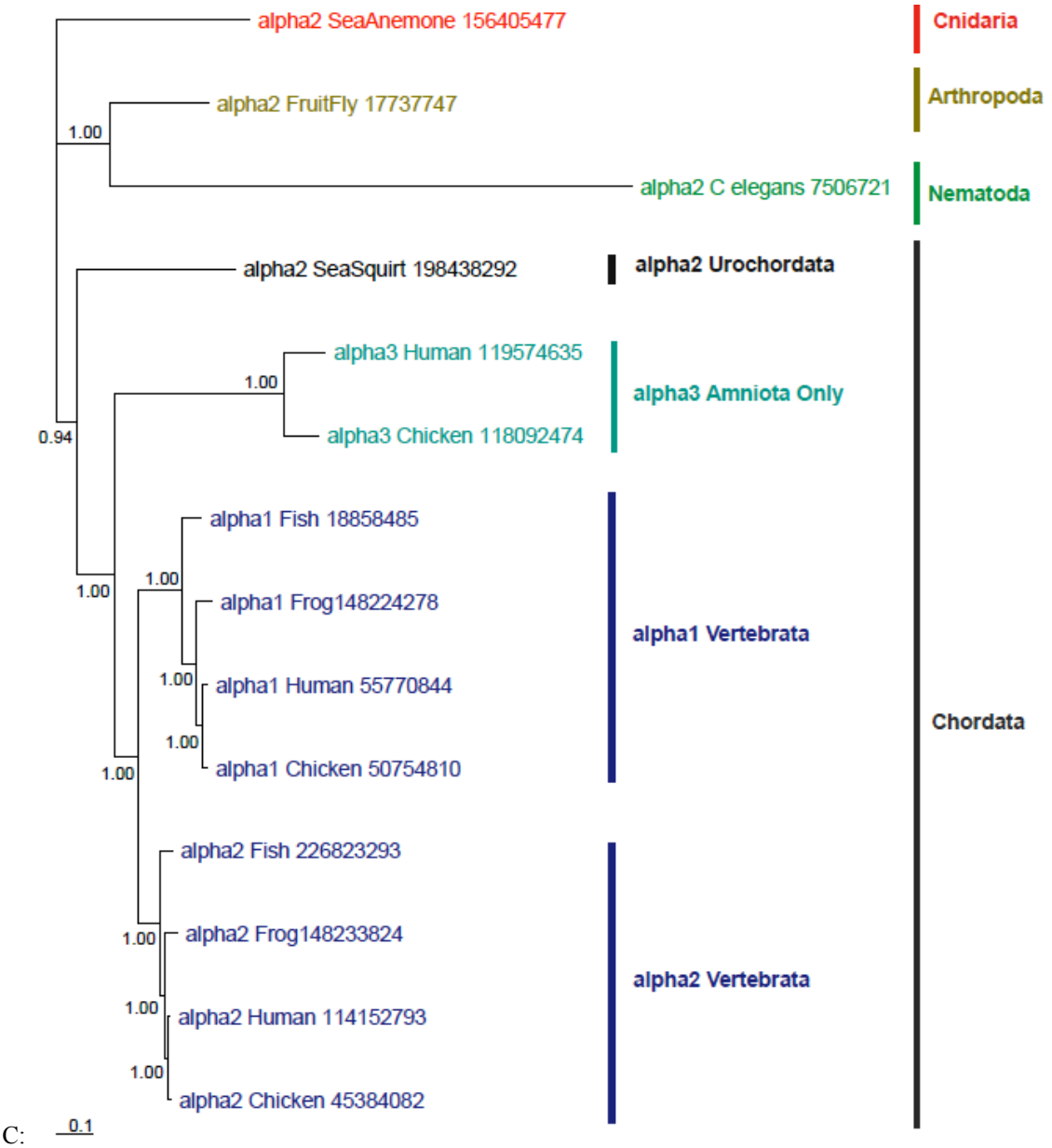
A: The Bayesian phylogeny for the p120 subfamily.

B: The Bayesian phylogeny for the beta catenin subfamily.

C: The Bayesian phylogeny for the alpha catenin subfamily.



Supplementary figure C.1 continued



C:

Supplementary figure C.1 continued

Supplementary table C.1: Summary of annotation problems for the catenin family.

'armp' stands for 'armadillo segment polarity protein'.

*Multiple GI numbers in the same box referred to the case that the same gene has been sequenced multiple times by different groups, and they are about 99-100% identical.

Phylum	Species	Sequence identifiers (NCBI GI number)*	Original annotation (gene symbol, gene description or definition)	Revised annotation (gene)
Arthropoda	*Drosophila melanogaster*	116007493; 113194561; 116007494; 116007496; 30923507; 6959880	p120	delta2/ARVCF
	Culex quinquefasciatus	170037762	Pkp4	delta2/ARVCF
	Nasonia vitripennis	156538289	p120	delta2/ARVCF
	Tribolium castaneum	189241994	similar to pkp4	delta2/ARVCF
	Pediculus humanus corporis	242010696	armadillo repeat protein, putative	delta2/ARVCF
	Acyrthosiphon pisum	193678745	similar to Adherens junction protein p120 CG17484-PB	delta2/ARVCF
Nematoda	*Caenorhabditis elegans*	75025630; 115532974; 285307422; 115532976; 87251663	jac-1	delta2/ARVCF
	Caenorhabditis briggsae	268534668	Cbr-jac-1	delta2/ARVCF
	Brugia malayi	170589179	Fibronectin type III domain containing protein	delta2/ARVCF
Platyhelminthes	*Schistosoma mansoni*	256078604	catenin and plakophilin	delta2/ARVCF
Cephalochordate	*Branchiostoma floridae*	260834528	hypothetical protein BRAFLDRAFT_146863	delta2/ARVCF
		260834522	hypothetical protein BRAFLDRAFT_246677	delta2/ARVCF
Chordata	*Xenopus laevis*	27447669	p120	pkp4
	Ciona intestinalis	198433544	similar to catenin, delta 1	ARVCF
	Danio rerio	190570248	armadillo repeat protein	ARVCF
Cnidaria	*Hydra magnipapillata*	221130487	similar to Adherens junction protein p120 CG17484-PB	delta2/ARVCF
		221131941	similar to predicted protein	delta2/ARVCF
	Nematostella vectensis	156379823	hypothetical protein	delta2/ARVCF
		156408524	hypothetical protein	delta2/ARVCF
Arthropoda	*Tribolium castaneum*	270014289	armadillo-1	beta catenin
		282165762	armadillo-2	beta catenin
	Pediculus humanus corporis	242012245	armp, putative	beta catenin
		242002974	armp, putative	beta catenin
	Acyrthosiphon pisum	193613160	similar to armadillo protein	beta catenin
		193613162	similar to armadillo protein	beta catenin

Supplementary table C.1 continued

Arthropoda	*Aedes aegypti*	122106728	armp	beta catenin
	Gryllus bimaculatus	37991668	armadillo protein	beta catenin
	Drosophila melanogaster	17136376; 45551205	armp;	beta catenin
	Drosophila pseudoobscura pseudoobscura	221222436; 198467818; 198146121	armp; Dpse\GA27602	beta catenin
	Drosophila yakuba	194187865; 195477916	Dyak\GE16998	beta catenin
	Culex quinquefasciatus	170040980	armadillo	beta catenin
	Nasonia vitripennis	156555532	armp	beta catenin
	Apis mellifera	297515465	armp	beta catenin
Nematoda	*Brugia malayi*	170590113	Armadillo/beta-catenin-like repeat family protein	hmp2 (beta catenin)
		170581705	Armadillo/beta-catenin-like repeat family protein	bar1 (beta catenin)
		170587340	Armadillo/beta-catenin-like repeat family protein	bar1 (beta catenin)
Lophotrochozoa	*Schistosoma mansoni*	256074627	plakoglobin	beta catenin
Cnidaria	*Nematostella vectensis*	156615300	hypothetical protein	beta catenin
Fugus	*Cryptococcus neoformans var. neoformans JEC21*	58258547	beta catenin	vac8p
Arthropoda	*Drosophila melanogaster*	17737747; 15291871	alpha catenin	alpha2 catenin
	Drosophila ananassae	194767509	Dana\GF20570	alpha2 catenin
	Drosophila erecta	194876539	Dere\GG16297	alpha2 catenin
	Drosophila grimshawi	195038657	alpha catenin	alpha2 catenin
	Drosophila mojavensis	195107690	Dmoj\GI23967	alpha2 catenin
	Drosophila persimilis	195151767	Dper\GL21877	alpha2 catenin
	Drosophila yakuba	195496883	Dyak\GE19475	alpha2 catenin
	Drosophila virilis	195400267	Dvir\GJ11153	alpha2 catenin
	Drosophila willistoni	195445209	Dwil\GK11940	alpha2 catenin
	Pediculus humanus corporis	242018616; 212514782	alpha1 catenin, putative	alpha2 catenin
	Ixodes scapularis	241730211	alpha catenin, putative	alpha2 catenin
	Acyrthosiphon pisum	193673870	similar to alpha Catenin CG17947-PA	alpha2 catenin
	Apis mellifera	66525427	alpha catenin	alpha2 catenin
	Tribolium castaneum	91076138	similar to actin binding	alpha2 catenin
	Aedes aegypti	157138056; 108880700; 157138058; 108880701	actin binding	alpha2 catenin
	Anopheles gambiae str. PEST	158290113	AgaP_AGAP003424	alpha2 catenin
	Culex quinquefasciatus	170038843	actin binding protein	alpha2 catenin
Cephalochordata	*Branchiostoma floridae*	260826764	hypothetical protein	alpha2 catenin
Echinodermata	*Strongylocentrotus purpuratus*	115894474	similar to alpha catenin	alpha2 catenin

Supplementary table C.1 continued

Echinodermata	*Lytechinus variegatus*	1098900	alpha catenin	alpha2 catenin
Hemichordata	*Saccoglossus kowalevskii*	268053957	alpha catenin	alpha2 catenin
Platyhelminthes	*Schistosoma mansoni*	256073504	alpha catenin	alpha2 catenin
Nematoda	*Caenorhabditis elegans*	193208515; 17563198; 74961297; 2738780; 7506721; 6434310	hmp-1; hypothetical protein R13H4.4; C. elegans protein R13H4.4a	alpha2 catenin
	Caenorhabditis briggsae	268557136; 187021170	Cbr-hmp-1	alpha2 catenin
	Brugia malayi	170573974	Vinculin family protein	alpha2 catenin
Cnidaria	*Nematostella vectensis*	156405477	hypothetical protein	alpha2 catenin

APPENDIX D

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

Supplementary Note D.1: Thioredoxin sequences used for ancestral sequences reconstruction.
GI numbers were accessed from GenBank. The names of the hosting organisms are also provided.

57164261 Ovis
27806783 Bos
47523692 Sus
126352340 Equus
6755911 Mus
16758644 Rattus
146291083 Rabit
135773 Human
67461921 Ponab
267126 Macmu
13560979 Callithrix
126339826 Monodelphis
149412981 Ornithorhynchus
45382053 Gallus
29373131 Melopsittacus
12958636 Ophiophagus
194332745 Xenopus
47215756 Tetraodon
9837585 Ictalurus
50539990 Danio
194160556 Drosophila
17648013 Drosophila
194141429 Drosophila
48104680 Apis
91084205 Tribolium
148298796 Bombyx
90819972 Graphocephala
169639275 Litopenaeus
30580603 Geocy
115401922 Aspergillus
119479067 Neosartorya
40746887 Aspergillus
115401518 Aspergillus
150951554 Pichia
46441186 Candida
126213085 Pichia
50309357 Kluyveromyces
151943486 Saccharomyces
50291653 Candida
151941211 Saccharomyces
19114764 Schizosaccharomyces
167537844 Monosiga
67479051 Entamoeba
165988451 Dictyostelium
15236327 Arabidopsis
15232567 Arabidopsis
154721452 Limonium
162461510 Zea
157335070 Vitis
145351136 Ostreococcus
53801490 Helicosporidium
1620905 Fagopyrum
46226985 Cryptosporidium
68350806 Theileria
148804689 Plasmodium
11498883 Archaeoglobus
116754023 Methanosaeta
91773622 Methanococcoides
154149646 Candidatus
88603734 Methanospirillum
48477193 Picrophilus
150401020 Methanococcus
124485138 Methanocorpusculum
116754438 Methanosaeta
76802488 Natronomonas
110667588 Haloquadratum
55380304 Haloarcula
76802694 Natronomonas
16120325 Halobacterium
11499727 Archaeoglobus
13541608 Thermoplasma
119720035 Thermofilum
159040636 Caldivirga
70607552 Sulfolobus
15899007 Sulfolobus
15922449 Sulfolobus
124027987 Hyperthermus
118431868 Aeropyrum
146304377 Metallosphaera
70607229 Sulfolobus
15897303 Sulfolobus
126465005 Staphylothermus

Supplementary Note D.1 continued

118431901 Aeropyrum
15894111 Clostridium
20808289 Thermoanaerobacter
16079205 Bacillus
16077522 Bacillus
15901736 Streptococcus
29377495 Enterococcus
153181008 Listeria
28377165 Lactobacillus
28379765 Lactobacillus
150393692 Staphylococcus
138896249 Geobacillus
30264587 Bacillus
16079902 Bacillus
28378864 Lactobacillus
153179313 Listeria
29375972 Enterococcus
15901605 Streptococcus
110798962 Clostridium
110800418 Clostridium
15894825 Clostridium
15896334 Clostridium
20807685 Thermoanaerobacter
76789276 Chlamydia
15836191 Chlamydophila
119357517 Chlorobium
119357012 Chlorobium
29345629 Bacteroides
150024368 Flavobacterium
34539910 Porphyromonas
29347639 Bacteroides
29346087 Bacteroides
34540117 Porphyromonas
29346866 Bacteroides
29345628 Bacteroides
32477354 Rhodopirellula
32476401 Rhodopirellula
15608608 Mycobacterium
57116870 Mycobacterium
62391823 Corynebacterium
72163169 Thermobifida
21219405 Streptomyces
72160576 Thermobifida
15607956 Mycobacterium
21219599 Streptomyces
62391938 Corynebacterium
21223797 Streptomyces
15611050 Mycobacterium
72163508 Thermobifida
21222296 Streptomyces
16329883 Synechocystis
17229833 Nostoc
17229385 Nostoc
16331440 Synechocystis
22299829 Thermosynechococcus
22297898 Thermosynechococcus
16329237 Synechocystis
22299630 Thermosynechococcus
17229697 Nostoc
17229859 Nostoc
22298354 Thermosynechococcus
17229358 Nostoc
126696505 Prochlorococcus
16331825 Synechocystis
17227548 Nostoc1
1351239 Pea Chloroplast
2507458 Spiol Chloroplast
11135474 Wheat Chloroplast
15594012 Pisum Chloroplast
11135407 Brana Chloroplast
46199419 Thermus
15807833 Deinococcus
46199687 Thermus
15805968 Deinococcus
147669275 Dehalococcoides
118047160 Chloroflexus
118048687 Chloroflexus
118046691 Chloroflexus
15606934 Aquifex
42521808 Bdellovibrio
39998535 Geobacter
42523902 Bdellovibrio
120602368 Desulfovibrio
39998370 Geobacter
116619824 Solibacter
116619449 Solibacter
94970094 Acidobacteria
34556879 Wolinella
15645443 Helicobacter
57237155 Campylobacter
15646067 Helicobacter
34557886 Wolinella

Supplementary Note D.1 continued

34556999 Wolinella
159184127 Agrobacterium
150398433 Sinorhizobium
17988305 Brucella
15603883 Rickettsia
108935910 Bovin Mitochondrio
194226778 Equus
21361403 Homo Mitochondrion
16758038 Rattus Mitochondrio
9903609 Mus Mitochondrion
74318624 Thiobacillus
121635072 Neisseria
74316054 Thiobacillus
126454139 Burkholderia
33602206 Bordetella
74318419 Thiobacillus
74316241 Thiobacillus
33602001 Bordetella
66043570 Pseudomonas
27364380 Vibrio
16124003 Yersinia
16767191 Salmonella
30064924 Shigella
67005950 Escherichia
16130507 Escherichia
30063983 Shigella
16765969 Salmonella
16123427 Yersinia
27366792 Vibrio

APPENDIX E

SUPPLEMENTARY INFORMATION FOR CHAPTER 6

Supplementary table E.1: List of 668 genes in the ancestor of the mycoides cluster (MCA – Mycoides Cluster Ancestor).

P$ The patterns for gene presence/absence in the five species MSB, MCAP, MSC, MLC, and Mf were consistent with the patterns in Figure 6.4B.

* The two genes (MCA_667 and MCA_668) in the MCA ancestor are unresolved (P17).

& All genes in the MCA ancestor are annotated with MLC, except five genes that are absent in MLC (P10, P8, and P11) are annotated with MCAP, and these five genes are MCA_640 to MCA_643, and MCA_663.

MCA	LocusTag	Gene product	Gene symbol	P$
MCA_1	MLC_4980	hypothetical protein	-	P1
MCA_2	MLC_0170	RpiR family transcriptional regulator	-	P1
MCA_3	MLC_2590	transmembrane protein	-	P1
MCA_4	MLC_7430	glucosamine 6 phosphate deaminase	nagB	P1
MCA_5	MLC_8750	putative membrane arginine transporter	arcD	P1
MCA_6	MLC_8840	hypothetical protein	-	P1
MCA_7	MLC_0390	hypothetical protein	-	P1
MCA_8	MLC_1080	transmembrane protein	-	P1
MCA_9	MLC_1830	putative oligo 1,6 glucosidase	dexA	P1
MCA_10	MLC_2400	oxidoreductase	-	P1
MCA_11	MLC_3850	transmembrane protein	-	P1
MCA_12	MLC_4640	hypothetical protein	-	P1
MCA_13	MLC_4960	glycerol transporter subunit C	gtsC	P1
MCA_14	MLC_5340	transcription repressor of the ROK family protein	-	P1
MCA_15	MLC_8340	hypothetical protein	-	P1
MCA_16	MLC_8820	amino acid permease	-	P1
MCA_17	MLC_8960	transmembrane protein	-	P1
MCA_18	MLC_0010	chromosomal replication initiator protein DnaA	dnaA	P1
MCA_19	MLC_0020	DNA polymerase III subunit beta	dnaN	P1
MCA_20	MLC_0030	Primase-like protein	-	P1
MCA_21	MLC_0040	Dimethyladenosine transferase	ksgA	P1
MCA_22	MLC_0060	DNA Gyrase Subunit B	gyrB	P1
MCA_23	MLC_0070	DNA Gyrase Subunit A	gyrA	P1
MCA_24	MLC_0080	ribose/galactose ABC transporter permease II	-	P1
MCA_25	MLC_0090	ribose/galactose ABC transporter permease I	-	P1
MCA_26	MLC_0100	ribose/galactose ABC transporter ATP-binding protein	-	P1
MCA_27	MLC_0120	Methionine tRNA synthetase	metG	P1
MCA_28	MLC_0220	30S ribosomal protein S18	S18	P1
MCA_29	MLC_0230	Single strand binding protein	ssb	P1
MCA_30	MLC_0240	30S ribosomal protein S6	rpsF	P1
MCA_31	MLC_0260	putative acyl carrier protein phosphodiesterase	-	P1
MCA_32	MLC_0270	ABC transporter ATP-binding protein	-	P1
MCA_33	MLC_0280	33 kDa chaperonin	hslO	P1
MCA_34	MLC_0360	ATP dependent zinc metallopeptidase FtsH	ftsH	P1
MCA_35	MLC_0380	tRNA(Ile) lysidine synthase	tilS	P1
MCA_36	MLC_0400	methyltransferase	-	P1
MCA_37	MLC_0410	DNA polymerase III subunit delta	holB	P1
MCA_38	MLC_0420	thymidylate kinase	tmk	P1
MCA_39	MLC_0430	Recombination protein	recR	P1
MCA_40	MLC_0440	DNA polymerase III subunits gamma/tau	dnaX	P1
MCA_41	MLC_0450	cytosine deaminase	codA	P1
MCA_42	MLC_0540	transmembrane protein	-	P1
MCA_43	MLC_0570	serine tRNA ligase	serS	P1
MCA_44	MLC_0580	hypothetical protein	-	P1

Supplementary table E.1 continued

MCA_45	MLC_0590	hypothetical protein	-	P1
MCA_46	MLC_0600	Lysine tRNA ligase	lysS	P1
MCA_47	MLC_0610	thioredoxin	trxA	P1
MCA_48	MLC_0620	hydrolase of the HAD family	-	P1
MCA_49	MLC_0630	transmembrane protein	-	P1
MCA_50	MLC_0640	phosphonate ABC transporter permease	phnE	P1
MCA_51	MLC_0650	phosphonate ABC transporter ATP-binding protein	phnC	P1
MCA_52	MLC_0660	phosphonate ABC transporter substrate-binding protein	-	P1
MCA_53	MLC_0670	asparagine tRNA ligase	asnS	P1
MCA_54	MLC_0680	hydrolase of the HAD family	-	P1
MCA_55	MLC_0690	hypothetical protein	-	P1
MCA_56	MLC_0700	O sialoglycoprotein endopeptidase	gcp	P1
MCA_57	MLC_0720	tRNA modification GTPase	trmE	P1
MCA_58	MLC_0730	30S ribosomal protein S20	rpsT	P1
MCA_59	MLC_0870	Preprotein translocase subunit SecA	secA	P1
MCA_60	MLC_0880	Proline dipeptidase	pepQ	P1
MCA_61	MLC_0890	DNA polymerase I, 5' 3' exonuclease	polA	P1
MCA_62	MLC_0940	hypothetical protein	-	P1
MCA_63	MLC_0950	dephospho CoA kinase	coaE	P1
MCA_64	MLC_0960	Hemolysin A	hlyA	P1
MCA_65	MLC_0970	hypothetical protein	-	P1
MCA_66	MLC_0980	exonuclease VII large subunit	xseA	P1
MCA_67	MLC_0990	transcription termination factor NusB	nusB	P1
MCA_68	MLC_1010	endonuclease IV	nfo	P1
MCA_69	MLC_1020	riboflavin kinase/FAD synthetase	ribC/ribF	P1
MCA_70	MLC_1090	transmembrane protein	-	P1
MCA_71	MLC_1100	ABC transporter ATP-binding protein	abc	P1
MCA_72	MLC_1180	glutamyl tRNA synthetase	gltX	P1
MCA_73	MLC_1200	DNA directed RNA polymerase subunit delta	rpoE	P1
MCA_74	MLC_1210	CTP synthase	pyrG	P1
MCA_75	MLC_1240	fructose bisphosphate aldolase class II	fbaA2	P1
MCA_76	MLC_1290	50S ribosomal protein L31	L31	P1
MCA_77	MLC_1300	hypothetical protein	-	P1
MCA_78	MLC_1310	phosphoesterase DHH family protein	dhh	P1
MCA_79	MLC_1320	thymidine kinase	tdk	P1
MCA_80	MLC_1330	peptide chain release factor 1	prfA	P1
MCA_81	MLC_1340	modification methylase	hemK	P1
MCA_82	MLC_1350	transmembrane protein	-	P1
MCA_83	MLC_1360	hypothetical protein	-	P1
MCA_84	MLC_1370	hypothetical protein	-	P1
MCA_85	MLC_1380	transmembrane protein	-	P1
MCA_86	MLC_1390	Cardiolipin synthetase	cls	P1
MCA_87	MLC_1400	30S ribosomal protein S12	rpsL	P1
MCA_88	MLC_1410	30S ribosomal protein S7	rpsG	P1
MCA_89	MLC_1420	Elongation factor G	fusA	P1
MCA_90	MLC_1430	Elongation factor Tu	tufA	P1
MCA_91	MLC_1450	alpha xylosidase or glucosidase	xylS	P1
MCA_92	MLC_1460	Leucyl aminopeptidase	pepA	P1
MCA_93	MLC_1470	hypothetical protein	-	P1
MCA_94	MLC_1480	tRNA (guanine N(7)) methyltransferase	trmB	P1
MCA_95	MLC_1490	Mg2+ transport protein	mgtE	P1
MCA_96	MLC_1550	alanyl tRNA synthetase	alaS	P1
MCA_97	MLC_1560	transmembrane protein	-	P1
MCA_98	MLC_1570	oligopeptide ABC transporter permease	oppB	P1
MCA_99	MLC_1580	oligopeptide ABC transporter permease I	oppC	P1
MCA_100	MLC_1590	oligopeptide ABC transporter ATP-binding protein	oppD	P1
MCA_101	MLC_1600	oligopeptide ABC transporter ATP-binding protein	oppF	P1
MCA_102	MLC_1610	oligopeptide ABC transporter substrate-binding protein	oppA	P1
MCA_103	MLC_1860	hypothetical protein	-	P1

Supplementary table E.1 continued

MCA_104	MLC_1920	spermidine/putrescine ABC transporter permease and substrate-binding protein	potCD	P1
MCA_105	MLC_1930	spermidine/putrescine ABC transporter permease	potB	P1
MCA_106	MLC_1940	spermidine/putrescine ABC transporter ATP-binding protein	potA	P1
MCA_107	MLC_1950	50S ribosomal protein L20	rplT	P1
MCA_108	MLC_1960	50s ribosomal protein L35	L35	P1
MCA_109	MLC_1970	translation initiation factor IF 3	infC	P1
MCA_110	MLC_1980	peptide deformylase 2	pdf	P1
MCA_111	MLC_1990	DNA methylase	-	P1
MCA_112	MLC_2000	guanylate kinase	gmk	P1
MCA_113	MLC_2010	Sun family protein	sun	P1
MCA_114	MLC_2040	GTP binding protein TypA/BipA	typA	P1
MCA_115	MLC_2300	endopeptidase	-	P1
MCA_116	MLC_2310	hypothetical protein	-	P1
MCA_117	MLC_2330	protein phosphatase	-	P1
MCA_118	MLC_2340	serine/threonine protein kinase	pkn	P1
MCA_119	MLC_2350	GTPase protein	-	P1
MCA_120	MLC_2360	Ribulose phosphate 3 epimerase	rpe	P1
MCA_121	MLC_2370	hypothetical protein	-	P1
MCA_122	MLC_2380	Valine tRNA ligase	valS	P1
MCA_123	MLC_2390	NAD kinase	-	P1
MCA_124	MLC_2410	hypothetical protein	-	P1
MCA_125	MLC_2430	transmembrane protein	-	P1
MCA_126	MLC_2440	excinuclease ABC subunit C	uvrC	P1
MCA_127	MLC_2450	transcription elongation factor	greA	P1
MCA_128	MLC_2460	oxidoreductase	-	P1
MCA_129	MLC_2470	hypothetical protein	-	P1
MCA_130	MLC_2480	hypothetical protein	-	P1
MCA_131	MLC_2510	putative GTP binding protein EngB	engB	P1
MCA_132	MLC_2520	Cation transporting ATPase	ctp	P1
MCA_133	MLC_2580	thiamin biosynthesis protein	thiI	P1
MCA_134	MLC_2600	30S ribosomal protein S4	rpsD	P1
MCA_135	MLC_2630	PTS system glucose specific transporter subunit IIABC	ptsG	P1
MCA_136	MLC_2640	phosphoenolpyruvate protein phosphotransferase	ptsI	P1
MCA_137	MLC_2670	acetate kinase	ackA	P1
MCA_138	MLC_2680	phosphate acetyltransferase	pta	P1
MCA_139	MLC_2690	dihydrolipoamide dehydrogenase	pdhD	P1
MCA_140	MLC_2700	dihydrolipoamide S acetyltransferase	pdhC	P1
MCA_141	MLC_2710	pyruvate dehydrogenase (lipoamide) subunit beta	pdhB	P1
MCA_142	MLC_2720	pyruvate dehydrogenase (lipoamide) subunit alpha	pdhA	P1
MCA_143	MLC_2730	lipoate protein ligase A	lplA	P1
MCA_144	MLC_2740	NADH oxidase	nox	P1
MCA_145	MLC_2760	Threonine tRNA ligase	thrS	P1
MCA_146	MLC_2770	pyruvate kinase	pyk	P1
MCA_147	MLC_2780	6 phosphofructokinase	pfk	P1
MCA_148	MLC_2820	hypoxanthine phosphoribosyltransferase	hpt	P1
MCA_149	MLC_2830	Holliday junction resolvase	-	P1
MCA_150	MLC_2840	hypothetical protein	-	P1
MCA_151	MLC_2850	2 phosphoglycerate dehydratase	eno	P1
MCA_152	MLC_3160	hypothetical protein	-	P1
MCA_153	MLC_3170	Proline tRNA ligase	proS	P1
MCA_154	MLC_3180	ribonuclease H-like protein	rnh	P1
MCA_155	MLC_3200	GTP binding protein LepA	lepA	P1
MCA_156	MLC_3210	hypothetical protein	-	P1
MCA_157	MLC_3220	aspartyl tRNA synthetase	aspS	P1
MCA_158	MLC_3230	histidyl tRNA synthetase	hisS	P1
MCA_159	MLC_3240	ribosome binding factor A	rbfA	P1
MCA_160	MLC_3250	tRNA pseudouridine synthase B	truB	P1
MCA_161	MLC_3260	riboflavin kinase	ribF	P1
MCA_162	MLC_3290	30S ribosomal protein S15	rpsO	P1
MCA_163	MLC_3300	transmembrane protein	-	P1

Supplementary table E.1 continued

MCA_164	MLC_3310	translation initiation factor IF 2	infB	P1
MCA_165	MLC_3320	hypothetical protein	-	P1
MCA_166	MLC_3330	hypothetical protein	-	P1
MCA_167	MLC_3340	transcriptional terminator	nusA	P1
MCA_168	MLC_3350	hypothetical protein	-	P1
MCA_169	MLC_3360	nitroreductase family protein	-	P1
MCA_170	MLC_3370	DNA polymerase III subunit alpha	polC	P1
MCA_171	MLC_3380	phosphatidate cytidylyltransferase	cdsA	P1
MCA_172	MLC_3390	Xaa Pro dipeptidase	pepQ	P1
MCA_173	MLC_3400	Tryptophan tRNA ligase	trpS	P1
MCA_174	MLC_3500	transketolase	tkt	P1
MCA_175	MLC_3510	transmembrane protein	-	P1
MCA_176	MLC_3650	transmembrane protein	-	P1
MCA_177	MLC_3660	transmembrane protein	-	P1
MCA_178	MLC_3670	hypothetical protein	-	P1
MCA_179	MLC_3680	hypothetical protein	-	P1
MCA_180	MLC_3690	23S rRNA pseudouridine synthase	rluB	P1
MCA_181	MLC_3700	deoxyguanosine kinase	dgk	P1
MCA_182	MLC_3720	transmembrane protein	-	P1
MCA_183	MLC_3760	hypothetical protein	-	P1
MCA_184	MLC_3780	ABC transporter ATP-binding protein	-	P1
MCA_185	MLC_3790	transmembrane protein	-	P1
MCA_186	MLC_3810	lipoprotein	-	P1
MCA_187	MLC_3830	inorganic pyrophosphatase	ppa	P1
MCA_188	MLC_3840	transmembrane protein	-	P1
MCA_189	MLC_3860	cytidylate kinase	cmk	P1
MCA_190	MLC_3870	GTP binding protein EngA	engA	P1
MCA_191	MLC_3880	glycerol 3 phosphate dehydrogenase [NAD(P)+]	gpsA	P1
MCA_192	MLC_3890	DNA binding protein HU	hup	P1
MCA_193	MLC_3900	hypothetical protein	-	P1
MCA_194	MLC_3910	Recombination protein U	recU	P1
MCA_195	MLC_3920	helicase	-	P1
MCA_196	MLC_3950	competence damage inducible protein	cinA	P1
MCA_197	MLC_3960	recombinase A	recA	P1
MCA_198	MLC_3970	hypothetical protein	-	P1
MCA_199	MLC_3980	signal recognition particle M54 protein	ffh	P1
MCA_200	MLC_3990	hypothetical protein	-	P1
MCA_201	MLC_4000	30S ribosomal protein S16	rpsP	P1
MCA_202	MLC_4010	16S rRNA processing protein rimM	rimM	P1
MCA_203	MLC_4020	tRNA (Guanine N(1)) methyltransferase	trmD	P1
MCA_204	MLC_4030	50S ribosomal protein L19	rplS	P1
MCA_205	MLC_4040	hypothetical protein	-	P1
MCA_206	MLC_4050	ribonuclease H II	rnhB	P1
MCA_207	MLC_4080	single strand binding protein	-	P1
MCA_208	MLC_4090	ABC transporter ATP-binding protein/permease	-	P1
MCA_209	MLC_4100	ABC transporter ATP-binding protein/permease	-	P1
MCA_210	MLC_4110	hypothetical protein	-	P1
MCA_211	MLC_4120	hypothetical protein	-	P1
MCA_212	MLC_4130	hypothetical protein	-	P1
MCA_213	MLC_4140	GTP binding protein Obg	obg	P1
MCA_214	MLC_4150	NH(3) dependent NAD(+) synthetase	nadE	P1
MCA_215	MLC_4170	putative nicotinate nucleotide adenylyltransferase	nadD	P1
MCA_216	MLC_4180	5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase	mtnN	P1
MCA_217	MLC_4240	tRNA(5 methylaminomethyl 2 thiouridylate) methyltransferase	trmU	P1
MCA_218	MLC_4250	transmembrane protein	-	P1
MCA_219	MLC_4290	hypothetical protein	-	P1
MCA_220	MLC_4300	methionyl tRNA formyltransferase	fmt	P1
MCA_221	MLC_4310	elongation factor P	efp	P1
MCA_222	MLC_4320	hypothetical protein	-	P1
MCA_223	MLC_4330	trigger factor Tig	tig	P1

Supplementary table E.1 continued

MCA_224	MLC_4340	ATP dependent protease La	lon	P1
MCA_225	MLC_4350	ATPase	-	P1
MCA_226	MLC_4390	transmembrane protein	-	P1
MCA_227	MLC_4400	hydrolase	-	P1
MCA_228	MLC_4410	ATP binding and permease	-	P1
MCA_229	MLC_4420	hypothetical protein	-	P1
MCA_230	MLC_4430	GTP binding protein Era	era	P1
MCA_231	MLC_4440	DNA repair protein RecO	recO	P1
MCA_232	MLC_4450	glycyl tRNA synthetase	glyS	P1
MCA_233	MLC_4460	DNA primase	dnaG	P1
MCA_234	MLC_4470	RNA polymerase sigma A factor	rpoD	P1
MCA_235	MLC_4480	hypothetical protein	-	P1
MCA_236	MLC_4490	hypothetical protein	-	P1
MCA_237	MLC_4500	ATP dependent RNA helicase	deaD	P1
MCA_238	MLC_4510	transmembrane protein	-	P1
MCA_239	MLC_4520	transmembrane protein	-	P1
MCA_240	MLC_4530	adenine phosphoribosyltransferase	apt	P1
MCA_241	MLC_4540	GTP pyrophosphokinase	relA	P1
MCA_242	MLC_4560	chromosome segregation ATPase	smc	P1
MCA_243	MLC_4590	ribonuclease III	rnc	P1
MCA_244	MLC_4600	Fatty acid/phospholipid synthesis protein PlsX	plsX	P1
MCA_245	MLC_4610	dihydroxyacetone related kinase	dhaK	P1
MCA_246	MLC_4620	hypothetical protein	-	P1
MCA_247	MLC_4630	50S ribosomal protein L28	L28	P1
MCA_248	MLC_4650	phosphate ABC transporter substrate-binding protein	pstS	P1
MCA_249	MLC_4660	phosphate ABC transporter permease	pstA	P1
MCA_250	MLC_4670	phosphate ABC transporter ATP-binding protein	pstB	P1
MCA_251	MLC_4680	phosphate transport system regulator PhoU	phoU	P1
MCA_252	MLC_4690	cell division protein FtsY	ftsY	P1
MCA_253	MLC_4700	hypothetical protein	-	P1
MCA_254	MLC_4710	hypothetical protein	-	P1
MCA_255	MLC_4720	S adenosylmethionine synthetase	metK	P1
MCA_256	MLC_4740	methylene tetrahydrofolate tRNA (uracil 5) methyltransferase	trmFO	P1
MCA_257	MLC_4760	Uracil DNA glycosylase	ung	P1
MCA_258	MLC_4770	CMP binding factor	cbf	P1
MCA_259	MLC_4780	histidine triad protein	hit	P1
MCA_260	MLC_4790	hypothetical protein	-	P1
MCA_261	MLC_4800	hypothetical protein	-	P1
MCA_262	MLC_4810	Nitrogen fixation protein NifS	nifS	P1
MCA_263	MLC_4820	Nitrogen fixation protein NifU	nifU	P1
MCA_264	MLC_4830	5 formyltetrahydrofolate cyclo ligase	-	P1
MCA_265	MLC_4840	endopeptidase O	pepO	P1
MCA_266	MLC_4850	glucose 6 phosphate isomerase	pgi	P1
MCA_267	MLC_4870	hypothetical protein	-	P1
MCA_268	MLC_4880	rRNA methylase	spoU	P1
MCA_269	MLC_4890	glyceraldehyde 3 phosphate dehydrogenase(NADP)	gapN	P1
MCA_270	MLC_4900	Topoisomerase IV subunit B	parE	P1
MCA_271	MLC_4910	Topoisomerase IV subunit A	parC	P1
MCA_272	MLC_4920	exodeoxyribonuclease V subunit alpha	recD	P1
MCA_273	MLC_4950	glycerol transporter subunit B	gtsB	P1
MCA_274	MLC_5130	L lactate dehydrogenase	ldh	P1
MCA_275	MLC_5150	transmembrane protein	-	P1
MCA_276	MLC_5180	hypothetical protein	-	P1
MCA_277	MLC_5210	30S ribosomal protein S21	rpsU	P1
MCA_278	MLC_5220	Holliday junction DNA helicase RuvA	ruvA	P1
MCA_279	MLC_5260	hypothetical protein	-	P1
MCA_280	MLC_5270	GTPase	-	P1
MCA_281	MLC_5280	DNA polymerase IV	dinB	P1
MCA_282	MLC_5300	Uridine kinase	udk	P1
MCA_283	MLC_5320	Xaa His dipeptidase	pepV	P1
MCA_284	MLC_5380	50S ribosomal protein L27	rpmA	P1

Supplementary table E.1 continued

MCA_285	MLC_5390	hypothetical protein	-	P1
MCA_286	MLC_5400	50s ribosomal protein L21	L21	P1
MCA_287	MLC_5410	hypothetical protein	-	P1
MCA_288	MLC_5420	tetrapyrrole methylase	-	P1
MCA_289	MLC_5440	hypothetical protein	-	P1
MCA_290	MLC_5460	1 acyl sn glycerol 3 phosphate acyltransferase	plsC	P1
MCA_291	MLC_5470	holo [acyl carrier protein]synthase	acpS	P1
MCA_292	MLC_5490	deoxycytidylate deaminase	dctD	P1
MCA_293	MLC_5500	transmembrane protein	-	P1
MCA_294	MLC_5510	ribosomal large subunit pseudouridylatesynthase D	rluD	P1
MCA_295	MLC_5520	lipoprotein signal peptidase	lsp	P1
MCA_296	MLC_5530	Isoleucine tRNA ligase	ileS	P1
MCA_297	MLC_5540	hypothetical protein	-	P1
MCA_298	MLC_5550	hypothetical protein	-	P1
MCA_299	MLC_5560	cell division protein FtsZ	ftsZ	P1
MCA_300	MLC_5570	hypothetical protein	-	P1
MCA_301	MLC_5580	S adenosyl methyltransferase MraW	mraW	P1
MCA_302	MLC_5590	hypothetical protein	-	P1
MCA_303	MLC_5600	50S ribosomal protein L32	L32	P1
MCA_304	MLC_5610	hypothetical protein	-	P1
MCA_305	MLC_5620	phenylalanyl tRNA synthetase beta chain	pheRS	P1
MCA_306	MLC_5630	phenylalanyl tRNA synthetase subunit alpha	pheS	P1
MCA_307	MLC_5640	hypothetical protein	-	P1
MCA_308	MLC_5650	ABC transporter permease	-	P1
MCA_309	MLC_5660	arginyl tRNA synthetase	argS	P1
MCA_310	MLC_5670	ribosome recycling factor	frr	P1
MCA_311	MLC_5680	Uridylate kinase	pyrH	P1
MCA_312	MLC_5690	transmembrane protein	-	P1
MCA_313	MLC_5700	Elongation factor Ts (EF Ts)	tsf	P1
MCA_314	MLC_5710	30S ribosomal protein S2	rpsB	P1
MCA_315	MLC_5720	chaperone protein DnaJ	dnaJ	P1
MCA_316	MLC_5730	chaperone protein DnaK	dnaK	P1
MCA_317	MLC_5740	GrpE protein (HSP 70 cofactor)	grpE	P1
MCA_318	MLC_5750	heat inducible transcription repressor HrcA	hrcA	P1
MCA_319	MLC_5760	ATP dependent Clp protease ATP binding subunit	clpB	P1
MCA_320	MLC_5780	tRNA/rRNA methyltransferase	spoU	P1
MCA_321	MLC_5790	HAM1-like protein	-	P1
MCA_322	MLC_6020	PTS system sucrose specifictransporter subunit IIBC	scrA	P1
MCA_323	MLC_6320	hypothetical protein	-	P1
MCA_324	MLC_6330	transmembrane protein	-	P1
MCA_325	MLC_6370	phosphoglycerate kinase	pgk	P1
MCA_326	MLC_6380	glyceraldehyde 3 phosphate dehydrogenase	gap	P1
MCA_327	MLC_6390	Primosomal protein DnaI	dnaI	P1
MCA_328	MLC_6400	chromosome replication initiation/membrane attachment protein	dnaB	P1
MCA_329	MLC_6410	formamidopyrimidine DNA glycosylase	fpg	P1
MCA_330	MLC_6420	DNA polymerase I	polA	P1
MCA_331	MLC_6430	DNA polymerase III subunit alpha	dnaE	P1
MCA_332	MLC_6440	Tyrosine tRNA ligase	tyrS	P1
MCA_333	MLC_6450	nicotinic phosphoribosyltransferase	pncB	P1
MCA_334	MLC_6460	phenylalanine tRNA ligase	pheT	P1
MCA_335	MLC_6470	hypothetical protein	-	P1
MCA_336	MLC_6480	hypothetical protein	-	P1
MCA_337	MLC_6490	ferric uptake regulator	fur	P1
MCA_338	MLC_6540	amino acid transporter	-	P1
MCA_339	MLC_6570	ABC transporter ATP-binding protein	-	P1
MCA_340	MLC_6600	transmembrane protein	-	P1
MCA_341	MLC_6610	Leucyl tRNA synthetase	leuS	P1
MCA_342	MLC_6630	30S ribosomal protein S9	rpsI	P1
MCA_343	MLC_6640	50S ribosomal protein L13	rplM	P1
MCA_344	MLC_6650	transmembrane protein permease	-	P1
MCA_345	MLC_6660	tRNA pseudouridine synthase A	truA	P1

Supplementary table E.1 continued

MCA_346	MLC_6670	ABC transporter permease	-	P1
MCA_347	MLC_6680	ABC transporter ATP-binding protein	-	P1
MCA_348	MLC_6690	ABC transporter ATP-binding protein	-	P1
MCA_349	MLC_6700	50S ribosomal protein L17	rplQ	P1
MCA_350	MLC_6710	DNA directed RNA polymerase subunit alpha	rpoA	P1
MCA_351	MLC_6720	30S ribosomal protein S11	rpsK	P1
MCA_352	MLC_6730	30S ribosomal protein S13	rpsM	P1
MCA_353	MLC_6740	translation initiation factor IF 1	infA	P1
MCA_354	MLC_6750	methionine aminopeptidase	map	P1
MCA_355	MLC_6760	adenylate kinase	adk	P1
MCA_356	MLC_6770	Preprotein translocase subunit secY	secY	P1
MCA_357	MLC_6780	50S ribosomal protein L15	rplO	P1
MCA_358	MLC_6790	30S ribosomal protein S5	rpsE	P1
MCA_359	MLC_6800	50S ribosomal protein L18	rplR	P1
MCA_360	MLC_6810	50S ribosomal protein L6	rplF	P1
MCA_361	MLC_6820	30S ribosomal protein S8	rpsH	P1
MCA_362	MLC_6840	50S ribosomal protein L5	rplE	P1
MCA_363	MLC_6850	50S ribosomal protein L24	L24	P1
MCA_364	MLC_6860	50S ribosomal protein L14	L14	P1
MCA_365	MLC_6870	30S ribosomal protein S17	S17	P1
MCA_366	MLC_6880	50S ribosomal protein L29	rpmC	P1
MCA_367	MLC_6890	50S ribosomal protein L16	rplP	P1
MCA_368	MLC_6900	30S ribosomal protein S3	rpsC	P1
MCA_369	MLC_6910	50S ribosomal protein L22	L22	P1
MCA_370	MLC_6915	30S ribosomal protein S19	rpsS	P1
MCA_371	MLC_6920	50S ribosomal protein L2	rplB	P1
MCA_372	MLC_6930	50S ribosomal protein L23	rplW	P1
MCA_373	MLC_6940	50S ribosomal protein L4	rplD	P1
MCA_374	MLC_6950	50S ribosomal protein L3	rplC	P1
MCA_375	MLC_6960	30S ribosomal protein S10	rpsJ	P1
MCA_376	MLC_7060	methylenetetrahydrofolate dehydrogenase	folD	P1
MCA_377	MLC_7070	K+, Na+ uptake protein bound cytoplasmic subunit	ktrB	P1
MCA_378	MLC_7080	hypothetical protein	-	P1
MCA_379	MLC_7090	aspartyl/glutamyl tRNA(Asn/Gln)amidotransferase subunit B	gatB	P1
MCA_380	MLC_7100	glutamyl tRNA(Gln) amidotransferase subunit A	gatA	P1
MCA_381	MLC_7110	aspartyl/glutamyl tRNA(Asn/Gln)amidotransferase subunit C	gatC	P1
MCA_382	MLC_7120	DNA ligase	ligA	P1
MCA_383	MLC_7130	transmembrane protein	-	P1
MCA_384	MLC_7140	RNA pseudouridylate synthase	rluC	P1
MCA_385	MLC_7150	CAAX amino terminalprotease family transmembrane protein	-	P1
MCA_386	MLC_7160	phosphocarrier protein HPr	ptsH	P1
MCA_387	MLC_7170	ATP dependent DNA helicase	pcrA	P1
MCA_388	MLC_7180	transmembrane protein	-	P1
MCA_389	MLC_7190	glycosyl transferase family protein	cps	P1
MCA_390	MLC_7210	peptide methionine sulfoxide reductase MsrA/MsrB	msrA	P1
MCA_391	MLC_7270	hypothetical protein	-	P1
MCA_392	MLC_7340	hydrolase of the HAD family	-	P1
MCA_393	MLC_7440	Triosephosphate isomerase	tpi	P1
MCA_394	MLC_7450	hydrolase of the HAD family	-	P1
MCA_395	MLC_7460	2,3 bisphosphoglycerate independent phosphoglycerate mutase	gpmI	P1
MCA_396	MLC_7490	deoxyribose phosphate aldolase	deoC	P1
MCA_397	MLC_7500	phosphoglucomutase or phosphomannomutase	manB	P1
MCA_398	MLC_7510	Thymidine phosphorylase	deoA	P1
MCA_399	MLC_7540	PTS system fructose specific transporter subunit IIBC	fruA	P1
MCA_400	MLC_7550	1 phosphofructokinase	fruB	P1
MCA_401	MLC_7560	transcription repressor of fructose operon	fruR	P1
MCA_402	MLC_7570	transmembrane protein	-	P1
MCA_403	MLC_7580	Purine nucleoside phosphorylase	pnp	P1

Supplementary table E.1 continued

MCA_404	MLC_7600	lysophospholipase	pldB	P1
MCA_405	MLC_7810	ribonucleoside diphosphate reductase subunit alpha	nrdE	P1
MCA_406	MLC_7820	ribonucleotide reductase	nrdI	P1
MCA_407	MLC_7830	ribonucleoside diphosphate reductase 2 betachain	nrdF	P1
MCA_408	MLC_7850	ribonuclease R	rnr	P1
MCA_409	MLC_7860	Ssra binding protein	smpB	P1
MCA_410	MLC_7890	PTS system glucose specific transporter subunit IIBC	ptsG	P1
MCA_411	MLC_7900	transmembrane protein	-	P1
MCA_412	MLC_7940	Mg(2+) transport ATPase, P type	mgtA	P1
MCA_413	MLC_7960	ATP synthase subunit epsilon	atpC	P1
MCA_414	MLC_7970	ATP synthase beta chain	atpD	P1
MCA_415	MLC_7980	ATP synthase subunit gamma	atpG	P1
MCA_416	MLC_7990	ATP synthase subunit alpha	atpA	P1
MCA_417	MLC_8000	ATP synthase subunit delta	atpH	P1
MCA_418	MLC_8010	ATP synthase subunit B	atpF	P1
MCA_419	MLC_8020	ATP synthase subunit C	atpE	P1
MCA_420	MLC_8030	ATP synthase subunit A	atpB	P1
MCA_421	MLC_8040	transmembrane protein	-	P1
MCA_422	MLC_8050	Uracil phosphoribosyltransferase	upp	P1
MCA_423	MLC_8060	serine hydroxymethyltransferase	glyA	P1
MCA_424	MLC_8070	ribose 5 phosphate isomerase RpiB	rpiB	P1
MCA_425	MLC_8080	rhodanese related sulfurtransferase	-	P1
MCA_426	MLC_8100	DNA directed RNA polymerase subunit beta'	rpoC	P1
MCA_427	MLC_8110	DNA directed RNA polymerase subunit beta	rpoB	P1
MCA_428	MLC_8130	50S ribosomal protein L7	L7	P1
MCA_429	MLC_8140	50S ribosomal protein L10	rplJ	P1
MCA_430	MLC_8150	50S ribosomal protein L1	rplA	P1
MCA_431	MLC_8160	50S ribosomal protein L11	rplK	P1
MCA_432	MLC_8280	transcription antitermination protein NusG	nusG	P1
MCA_433	MLC_8290	Preprotein translocase subunit SecE	secE	P1
MCA_434	MLC_8310	tRNA/rRNA methyltransferase	spoU	P1
MCA_435	MLC_8320	cysteinyl tRNA synthetase	cysS	P1
MCA_436	MLC_8330	transmembrane protein	-	P1
MCA_437	MLC_8350	Replicative DNA helicase DnaC	dnaC	P1
MCA_438	MLC_8360	50S ribosomal protein L9	rplI	P1
MCA_439	MLC_8370	peptidyl tRNA hydrolase	pth	P1
MCA_440	MLC_8380	ribose phosphate pyrophosphokinase	prs	P1
MCA_441	MLC_8400	TatD related deoxyribonuclease	-	P1
MCA_442	MLC_8410	transmembrane protein	-	P1
MCA_443	MLC_8420	hypothetical protein	-	P1
MCA_444	MLC_8430	excinuclease ABC subunit B	uvrB	P1
MCA_445	MLC_8440	excinuclease ABC subunit A	uvrA	P1
MCA_446	MLC_8450	dihydrofolate:folylpolyglutamate synthase	folC	P1
MCA_447	MLC_8460	transmembrane protein	-	P1
MCA_448	MLC_8470	HPr kinase/phosphorylase	hprK	P1
MCA_449	MLC_8480	prolipoprotein diacylglyceryl transferase	lgt	P1
MCA_450	MLC_8490	thioredoxin reductase	trxB	P1
MCA_451	MLC_8500	prolipoprotein diacylglyceryl transferase	lgt	P1
MCA_452	MLC_8510	hypothetical protein	-	P1
MCA_453	MLC_8620	DNA topoisomerase I	topA	P1
MCA_454	MLC_8760	hypothetical protein	-	P1
MCA_455	MLC_8770	GTP binding protein	engD	P1
MCA_456	MLC_8780	methyltransferase GidB	gidB	P1
MCA_457	MLC_8790	CDP diacylglycerol glycerol 3 phosphate 3 phosphatidyltransferase	pgsA	P1
MCA_458	MLC_8810	transmembrane protein	-	P1
MCA_459	MLC_8890	tRNA uridine 5 carboxymethylaminomethyl modification enzyme	mnmG	P1
MCA_460	MLC_8900	hypothetical protein	-	P1
MCA_461	MLC_8910	NADH oxidase	nox	P1
MCA_462	MLC_8920	Pyrazinamidase/nicotinamidase	pncA	P1

Supplementary table E.1 continued

MCA_463	MLC_9190	hypothetical protein	-	P1
MCA_464	MLC_9200	membrane protein OxaA	oxaA	P1
MCA_465	MLC_9210	ribonuclease P protein component	rnpA	P1
MCA_466	MLC_9220	50S ribosomal protein L34	rpmH	P1
MCA_467	MLC_2960	hypothetical protein	-	P2
MCA_468	MLC_8660	transmembrane protein	-	P2
MCA_469	MLC_8680	lipoprotein	-	P2
MCA_470	MLC_1040	glycosyltransferase	epsG	P2
MCA_471	MLC_1500	hypothetical protein	-	P2
MCA_472	MLC_5480	transmembrane protein	-	P2
MCA_473	MLC_5930	transmembrane protein	-	P2
MCA_474	MLC_7020	hypothetical protein	-	P2
MCA_475	MLC_3460	transmembrane protein	-	P2
MCA_476	MLC_7410	hypothetical protein	-	P2
MCA_477	MLC_1230	hypothetical protein	-	P2
MCA_478	MLC_5950	transmembrane protein	-	P2
MCA_479	MLC_7010	hypothetical protein	-	P2
MCA_480	MLC_0050	transmembrane protein	-	P2
MCA_481	MLC_0140	hypothetical protein	-	P2
MCA_482	MLC_0160	mannitol 1 phosphate 5 dehydrogenase	mtlD	P2
MCA_483	MLC_0180	PTS system mannitol transporter subunit IIA	mtlF	P2
MCA_484	MLC_0190	Sorbitol 6 phosphate 2 dehydrogenase	srlD	P2
MCA_485	MLC_0200	PTS system mannitol transporter subunit IIBC	mtlA	P2
MCA_486	MLC_0210	putative DNA recombinase RecG	recG	P2
MCA_487	MLC_0250	Cold shock protein	csp	P2
MCA_488	MLC_0290	transmembrane protein	-	P2
MCA_489	MLC_0310	transmembrane protein permease	-	P2
MCA_490	MLC_0370	hypothetical protein	-	P2
MCA_491	MLC_0470	transmembrane protein	-	P2
MCA_492	MLC_0490	hypothetical protein	-	P2
MCA_493	MLC_0500	putative peroxiredoxin Bcp	bcp	P2
MCA_494	MLC_0530	transmembrane protein	-	P2
MCA_495	MLC_0560	transmembrane protein	-	P2
MCA_496	MLC_0710	hypothetical protein	-	P2
MCA_497	MLC_0850	transmembrane protein	-	P2
MCA_498	MLC_0860	transmembrane protein	-	P2
MCA_499	MLC_1000	hypothetical protein	-	P2
MCA_500	MLC_1120	Threonine dehydratase	ilvA	P2
MCA_501	MLC_1130	transmembrane protein	-	P2
MCA_502	MLC_1140	transmembrane protein	-	P2
MCA_503	MLC_1170	PTS system N acetylglucosamine specific transporter subunit IIBC	nagE	P2
MCA_504	MLC_1250	AAA family ATPase	-	P2
MCA_505	MLC_1270	transmembrane protein permease	-	P2
MCA_506	MLC_1280	glycerophosphoryl diester phosphodiesterase	glpQ	P2
MCA_507	MLC_1440	PTS system transporter subunit IIBC	PtsG	P2
MCA_508	MLC_1850	oligoendopeptidase F	pepF	P2
MCA_509	MLC_1880	hypothetical protein	-	P2
MCA_510	MLC_1890	chromate transport protein	-	P2
MCA_511	MLC_2320	transmembrane protein	-	P2
MCA_512	MLC_2490	transmembrane protein	-	P2
MCA_513	MLC_2500	transmembrane protein	-	P2
MCA_514	MLC_2530	transmembrane protein	-	P2
MCA_515	MLC_2540	hypothetical protein	-	P2
MCA_516	MLC_2560	transmembrane protein	-	P2
MCA_517	MLC_2570	transmembrane protein and tail specific protease	-	P2
MCA_518	MLC_2610	hypothetical protein	-	P2
MCA_519	MLC_2620	glycerone kinase	dhaK2	P2
MCA_520	MLC_2650	putative phosphopantetheine adenylyltransferase	coaD	P2
MCA_521	MLC_2660	Prolipoprotein	-	P2
MCA_522	MLC_2790	glycerol 3 phosphate oxidase	glpO	P2
MCA_523	MLC_2800	glycerol kinase	glpK	P2
MCA_524	MLC_2810	glycerol facilitator factor	glpF	P2

Supplementary table E.1 continued

MCA_525	MLC_3090	transmembrane protein	-	P2
MCA_526	MLC_3110	hypothetical protein	-	P2
MCA_527	MLC_3190	phospholipase	-	P2
MCA_528	MLC_3270	transmembrane protein protease	-	P2
MCA_529	MLC_3280	hypothetical protein	-	P2
MCA_530	MLC_3410	transmembrane protein	-	P2
MCA_531	MLC_3450	DNA recombination protein RmuC	-	P2
MCA_532	MLC_3470	methylenetetrahydrofolate tRNA (uracil 5) methyltransferase	trmFO	P2
MCA_533	MLC_3480	transmembrane protein	-	P2
MCA_534	MLC_3490	hypothetical protein	-	P2
MCA_535	MLC_3520	hypothetical protein	-	P2
MCA_536	MLC_3750	PTS system transporter subunit IIA	pts	P2
MCA_537	MLC_4160	transmembrane protein	-	P2
MCA_538	MLC_4190	deoxynucleoside kinase	-	P2
MCA_539	MLC_4570	hypothetical protein	-	P2
MCA_540	MLC_4580	hypothetical protein	-	P2
MCA_541	MLC_4730	Copper homeostasis protein	cut	P2
MCA_542	MLC_4750	mannose 6 phosphate isomerase	pmi	P2
MCA_543	MLC_4860	transmembrane protein	-	P2
MCA_544	MLC_4930	transmembrane protein permease	-	P2
MCA_545	MLC_4940	glycerol transporter subunit A	gtsA	P2
MCA_546	MLC_4970	lipoprotein	-	P2
MCA_547	MLC_5010	NADH dependent flavin oxidoreductase	-	P2
MCA_548	MLC_5020	lipoate protein ligase A	lplA	P2
MCA_549	MLC_5030	hypothetical protein	-	P2
MCA_550	MLC_5040	glycine cleavage system H protein	gcdH	P2
MCA_551	MLC_5050	triacylglycerol lipase	lip	P2
MCA_552	MLC_5070	triacylglycerol lipase	lip	P2
MCA_553	MLC_5140	N acetylglucosamine 6 phosphate deacetylase	nagA	P2
MCA_554	MLC_5160	transmembrane protein	-	P2
MCA_555	MLC_5170	transmembrane protein	-	P2
MCA_556	MLC_5240	dihydrolipoyl dehydrogenase	pdhD	P2
MCA_557	MLC_5250	hypothetical protein	-	P2
MCA_558	MLC_5310	hypothetical protein	-	P2
MCA_559	MLC_5330	N acetylmannosamine 6 phosphate 2 epimerase	nanE	P2
MCA_560	MLC_5350	hypothetical protein	-	P2
MCA_561	MLC_5360	sodium:solute symporter family	-	P2
MCA_562	MLC_5370	N acetylneuraminate lyase	nanA	P2
MCA_563	MLC_5450	hypothetical protein	-	P2
MCA_564	MLC_5770	hydrolase of the HAD family	-	P2
MCA_565	MLC_5820	hypothetical protein	-	P2
MCA_566	MLC_5830	ATP synthase beta chain	atpD	P2
MCA_567	MLC_5840	ATP synthase subunit alpha	atpA	P2
MCA_568	MLC_5850	transmembrane protein	-	P2
MCA_569	MLC_5860	hypothetical protein	-	P2
MCA_570	MLC_5870	hypothetical protein	-	P2
MCA_571	MLC_5880	hypothetical protein	-	P2
MCA_572	MLC_5890	transmembrane protein	-	P2
MCA_573	MLC_5900	lipoprotein	-	P2
MCA_574	MLC_5910	transmembrane protein	-	P2
MCA_575	MLC_5920	hypothetical protein	-	P2
MCA_576	MLC_5960	hypothetical protein	-	P2
MCA_577	MLC_6000	fructose bisphosphate aldolase class II	fba	P2
MCA_578	MLC_6010	transcriptional repressor	fruR	P2
MCA_579	MLC_6030	1 phosphofructokinase	fruK	P2
MCA_580	MLC_6170	hypothetical protein	-	P2
MCA_581	MLC_6180	transmembrane protein	-	P2
MCA_582	MLC_6190	hypothetical protein	-	P2
MCA_583	MLC_6500	hypothetical protein	-	P2
MCA_584	MLC_6510	transmembrane protein	-	P2
MCA_585	MLC_6520	carbamate kinase	arcC	P2
MCA_586	MLC_6530	putative agmatine deiminase	aguA	P2

Supplementary table E.1 continued

MCA_587	MLC_6550	ornithine carbamoyltransferase	arcB	P2
MCA_588	MLC_6580	transmembrane protein	-	P2
MCA_589	MLC_6590	transmembrane protein	-	P2
MCA_590	MLC_6620	hypothetical protein	-	P2
MCA_591	MLC_6970	glycerone kinase	dhaK1	P2
MCA_592	MLC_7050	hypothetical protein	-	P2
MCA_593	MLC_7220	hypothetical protein	-	P2
MCA_594	MLC_7240	transmembrane protein	-	P2
MCA_595	MLC_7280	hypothetical protein	-	P2
MCA_596	MLC_7290	putative C5 methylase	marMP	P2
MCA_597	MLC_7310	alkylphosphonate ABC transporter permease	phnB	P2
MCA_598	MLC_7320	alkylphosphonate ABC transporter ATP-binding protein	phnC	P2
MCA_599	MLC_7330	alkylphosphonate ABC transporter substrate-binding protein	phnD	P2
MCA_600	MLC_7350	aminotransferase	patB	P2
MCA_601	MLC_7360	hypothetical protein	-	P2
MCA_602	MLC_7420	hypothetical protein	-	P2
MCA_603	MLC_7470	hypothetical protein	-	P2
MCA_604	MLC_7480	hypothetical protein	-	P2
MCA_605	MLC_7590	putative new IS transposase protein A	tnp	P2
MCA_606	MLC_7630	Fic family protein	-	P2
MCA_607	MLC_7870	hypothetical protein	-	P2
MCA_608	MLC_7950	hypothetical protein	-	P2
MCA_609	MLC_8090	hypothetical protein	-	P2
MCA_610	MLC_8170	UTP glucose 1 phosphate uridylyltransferase	galU	P2
MCA_611	MLC_8220	oligopeptide ABC transporter ATP-binding protein	oppF	P2
MCA_612	MLC_8230	oligopeptide ABC transporter ATP-binding protein	oppD	P2
MCA_613	MLC_8240	oligopeptide ABC transporter permease	oppC	P2
MCA_614	MLC_8250	oligopeptide ABC transporter permease	oppB	P2
MCA_615	MLC_8260	oligopeptide ABC transporter substrate-binding protein	oppA	P2
MCA_616	MLC_8270	transmembrane protein	-	P2
MCA_617	MLC_8610	hypothetical protein	-	P2
MCA_618	MLC_8690	ABC transporter ATP-binding protein	abc	P2
MCA_619	MLC_8710	hypothetical protein	-	P2
MCA_620	MLC_8720	membrane protein	-	P2
MCA_621	MLC_8730	lipoprotein	-	P2
MCA_622	MLC_8740	transmembrane protein	-	P2
MCA_623	MLC_8830	magnesium transporting ATPase, P type 1	mgtA	P2
MCA_624	MLC_8850	Hexose phosphate transport protein	uhpT	P2
MCA_625	MLC_8860	hypothetical protein	-	P2
MCA_626	MLC_8870	aspartate ammonia ligase	asnA	P2
MCA_627	MLC_9180	PTS system lichenan specific transporter subunit IIA	licA	P2
MCA_628	MLC_1110	TatD related deoxyribonuclease	-	P3
MCA_629	MLC_1840	transcriptional regulator	-	P3
MCA_630	MLC_3710	hypothetical protein	-	P3
MCA_631	MLC_6735	50S ribosomal protein L36	rpmJ	P3
MCA_632	MLC_6830	30S ribosomal protein S14	rpsN	P3
MCA_633	MLC_7840	Preprotein translocase subunit SecG	secG	P3
MCA_634	MLC_8120	hypothetical protein	-	P3
MCA_635	MLC_8300	50S ribosomal protein L33	L33	P3
MCA_636	MLC_5120	transmembrane protein	-	P4
MCA_637	MLC_3550	hypothetical protein	-	P5
MCA_638	MLC_7740	adenylosuccinate lyase	purB	P5
MCA_639	MLC_7750	adenylosuccinate synthetase	purA	P5
MCA_640&	MCAP_0284	hypothetical protein	-	P10
MCA_641&	MCAP_0647	acyl carrier protein	-	P8
MCA_642&	MCAP_0806	hypothetical protein	-	P8
MCA_643&	MCAP_0589	ribulose-phosphate 3-epimerase	-	P8
MCA_644	MLC_0460	GMP reductase	guaC	P14
MCA_645	MLC_0830	hypothetical protein	-	P15

Supplementary table E.1 continued

MCA_646	MLC_1800	maltose ABC transporter permease	malG	P15
MCA_647	MLC_0110	ribose/galactose ABC transporter substrate-binding protein	-	P13
MCA_648	MLC_0340	malate permease	-	P13
MCA_649	MLC_1190	hypothetical protein	-	P13
MCA_650	MLC_2020	cytosine specific DNA methyltransferase Sau96I	dcm	P13
MCA_651	MLC_2030	type II site specific deoxyribonuclease	sau96I -like	P13
MCA_652	MLC_7000	transcriptional regulator	treR	P13
MCA_653	MLC_7670	beta glucosidase	bgl	P13
MCA_654	MLC_1670	hypothetical protein	-	P13
MCA_655	MLC_2860	GntR family transcriptional regulator	-	P13
MCA_656	MLC_4380	hypothetical protein	-	P13
MCA_657	MLC_5230	Holliday junction DNA helicase RuvB	ruvB	P13
MCA_658	MLC_6980	alpha amylase	treA	P13
MCA_659	MLC_6990	PTS system trehalose specific transporter subunit IIBC	treP	P13
MCA_660	MLC_7680	sugar kinase, ROK family	suk	P13
MCA_661	MLC_7910	glucokinase	glk	P13
MCA_662	MLC_8800	aminoacid permease	-	P13
MCA_663&	MCAP_0186	hypothetical protein	-	P11
MCA_664	MLC_1810	maltodextrin ABC transporter ATP-binding protein	malK	P7
MCA_665	MLC_7530	IS1296 H transposase protein B	tnp	P6
MCA_666	MLC_0520	putative adenine specific DNA methyltransferase	dam	P17
MCA_667*	MLC_0750	IS1296 A transposase protein B	tnp	P17*
MCA_668*	MLC_2870	PTS system sucrose specific transporter subunit IIBC	scrA	P17*

Supplementary table E.2: List of 41 genes present in the mycoides cluster ancestor (MCA) and the proteome of *M. genitalium* (Mg476), but absent in the minimal genome of *M. genitalium* (Mg381).

MG	gene product	gene symbol	Length(aa)	MCA
MG_009	deoxyribonuclease, TatD family, putative	-	262	MCA_441
MG_018	helicase SNF2 family, putative	-	1031	MCA_535
MG_033	glycerol uptake facilitator	glpF	258	MCA_524
MG_039	FAD-dependent glycerol-3-phosphate dehydrogenase, putative	-	384	MCA_522
MG_051	pyrimidine-nucleoside phosphorylase	pdp	421	MCA_398
MG_056	tetrapyrrole (corrin/porphyrin) methylase protein	-	277	MCA_288
MG_061	Mycoplasma MFS transporter	-	567	MCA_624
MG_062	PTS system, fructose-specific IIABC component	fruA	680	MCA_399
MG_063	1-phosphofructokinase, putative	fruK	303	MCA_400
MG_066	transketolase	tkt	648	MCA_174
MG_103	conserved hypothetical protein	-	280	MCA_452
MG_110	ribosome small subunit-dependent GTPase A	rsgA	278	MCA_119
MG_112	ribulose-phosphate 3-epimerase	rpe	209	MCA_120
MG_114	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	-	236	MCA_457
MG_115	competence/damage-inducible protein CinA domain protein	-	157	MCA_196
MG_121	ABC transporter, permease protein	-	306	MCA_24
MG_183	oligoendopeptidase F	-	607	MCA_508
MG_210	signal peptidase II	-	181	MCA_295
MG_214	segregation and condensation protein B	-	207	MCA_179
MG_244	UvrD/REP helicase	-	703	MCA_387
MG_248	conserved hypothetical protein	-	218	MCA_235
MG_252	RNA methyltransferase, TrmH family, group 3	-	242	MCA_434
MG_498	formamidopyrimidine-DNA glycosylase	mutM	284	MCA_329
MG_289	phosphonate ABC transporter, substrate binding protein (P37), putative	-	368	MCA_599
MG_291	phosphonate ABC transporter, permease protein (P69), putative	-	543	MCA_597
MG_293	glycerophosphoryl diester phosphodiesterase family protein	-	244	MCA_506
MG_294	major facilitator superfamily protein, putative	-	474	MCA_505
MG_298	chromosome segregation protein SMC	smc	982	MCA_242
MG_339	recA protein (recombinase A)	recA	340	MCA_197
MG_346	RNA methyltransferase, TrmH family, group 2	-	166	MCA_320
MG_352	recombination protein U	recU	166	MCA_194
MG_360	ImpB/MucB/SamB family protein	-	411	MCA_281
MG_367	ribonuclease III	rnc	262	MCA_243
MG_380	methyltransferase GidB	-	192	MCA_456
MG_390	ABC transporter, ATP-binding/permease protein	-	660	MCA_228
MG_398	ATP synthase F1, epsilon subunit	atpC	133	MCA_413
MG_411	phosphate ABC transporter, permease protein PstA	-	654	MCA_249
MG_428	LuxR bacterial regulatory protein, putative	-	171	MCA_93
MG_437	phosphatidate cytidylyltransferase	cdsA	397	MCA_171
MG_460	L-lactate dehydrogenase/malate dehydrogenase	ldh	312	MCA_274
MG_463	dimethyladenosine transferase	-	259	MCA_21

Supplementary table E.3: List of 105 genes present in MCA and the proteome of M. pulmonis, but absent in the minimal genome of M. pulmonis (Mp310).

MG	gene product	gene symbol	Length (aa)	MCA
MYPU_0140	conserved hypothetical protein	-	252	MCA_441
MYPU_0150	DIMETHYLADENOSINE TRANSFERASE (S-ADENOSYLMETHIONINE-6-N', N'-ADENOSYL(RRNA) DIMETHYLTRANSFERASE)(16S RRNA DIMETHYLASE)(HIGH LEVEL KASUGAMYCIN RESISTANCE PROTEIN KSGA) (KASUGAMYCIN DIMETHYLTRANSFERASE)	ksgA	252	MCA_21
MYPU_0170	PTS SYSTEM, GLUCOSE-SPECIFIC IIABC COMPONENT (EIIABC-GLC) (GLUCOSE-PERMEASE IIABC COMPONENT) (PHOSPHOTRANSFERASE ENZYME II, ABC COMPONENT)(EII-GLC/EIII-GLC)	ptsG	791	MCA_135
MYPU_0320	LIPOATE-PROTEIN LIGASE A	lplA	345	MCA_143
MYPU_0350	ESTERASE/LIPASE 1	-	268	MCA_552
MYPU_0430	CPG DNA METHYLASE (CYTOSINE-SPECIFIC METHYLTRANSFERASE)	-	296	MCA_650
MYPU_0510	RECOMBINATION PROTEIN RECR	recR	191	MCA_39
MYPU_0540	conserved hypothetical protein	-	222	MCA_288
MYPU_0690	ATP-DEPENDENT RNA HELICASE	deaD	480	MCA_237
MYPU_0860	conserved hypothetical protein	-	157	MCA_194
MYPU_0960	EXCINUCLEASE ABC SUBUNIT B	uvrB	657	MCA_444
MYPU_1280	GLUCOSE INHIBITED DIVISION PROTEIN B	gidB	227	MCA_456
MYPU_1290	PTS SYSTEM, FRUCTOSE-SPECIFIC IIABC COMPONENT (EIIABC-FRU) (FRUCTOSE-PERMEASE IIABC COMPONENT)(PHOSPHOTRANSFERASE ENZYME II, ABC COMPONENT) (EII-FRU/EIII-FRU)	fruA	698	MCA_399
MYPU_1450	THYMIDINE KINASE	tdk	190	MCA_79
MYPU_1560	EXCINUCLEASE ABC SUBUNIT C	uvrC	563	MCA_126
MYPU_1590	unknown; predicted coding region	-	228	MCA_231
MYPU_1630	RIBONUCLEASE III (RNASE III)	rnc	241	MCA_243
MYPU_1660	conserved hypothetical protein	-	203	MCA_94
MYPU_1690	unknown; predicted coding region	-	160	MCA_67
MYPU_1710	HEMOLYSIN A	-	238	MCA_64
MYPU_1780	TRNA/RRNA METHYLTRANSFERASE	-	229	MCA_434
MYPU_1870	conserved hypothetical protein	-	963	MCA_595
MYPU_1880	DNA-DAMAGE REPAIR PROTEIN MUCB	mucB	425	MCA_281
MYPU_1930	LIPOPROTEIN	-	323	MCA_468
MYPU_1940	LIPOPROTEIN	-	280	MCA_620
MYPU_1950	LIPOPROTEIN	-	320	MCA_621
MYPU_1970	unknown; predicted coding region	-	606	MCA_622
MYPU_1980	TRANSPORT PROTEIN SGAT, LIPOPROTEIN	sgaT	651	MCA_183
MYPU_2040	conserved hypothetical protein	-	103	MCA_166
MYPU_2090	TRIGGER FACTOR (PROLYL ISOMERASE)	tig	459	MCA_223
MYPU_2180	P69-LIKE (Mycoplasma hyorhinis) ABC TRANSPORTER PERMEASE PROTEIN	-	581	MCA_597
MYPU_2210	GLYCEROL KINASE (ATP:GLYCEROL 3-PHOSPHOTRANSFERASE) (GLYCEROKINASE) (GK)	glpK	507	MCA_523
MYPU_2220	GLYCEROL UPTAKE FACILITATOR PROTEIN	glpF	247	MCA_524
MYPU_2260	conserved hypothetical protein	-	287	MCA_452
MYPU_2280	GLUCOKINASE (GLUCOSE KINASE)	glcK	287	MCA_14
MYPU_2520	RECOMBINATION PROTEIN RECA	recA	339	MCA_197
MYPU_2540	conserved hypothetical protein	-	152	MCA_200

Supplementary table E.3 continued

MYPU_2550	conserved hypothetical protein	-	324	MCA_436
MYPU_2620	conserved hypothetical protein	-	467	MCA_544
MYPU_2630	GLYCEROPHOSPHORYL DIESTER PHOSPHODIESTERASE (GLYCEROPHOSPHODIESTER PHOSPHODIESTERASE)	glpQ	240	MCA_506
MYPU_2640	GLYCEROL-3-PHOSPHATE DEHYDROGENASE (G-3-P DEHYDROGENASE)	glpD	384	MCA_522
MYPU_2650	ATP SYNTHASE EPSILON CHAIN	atpC	141	MCA_413
MYPU_2760	PROLINE DIPEPTIDASE PEPQ	pepQ	120	MCA_60
MYPU_2780	DEOXYCYTIDYLATE DEAMINASE (DCMP DEAMINASE)	dctD	154	MCA_292
MYPU_2880	LIPOPROTEIN	-	785	MCA_521
MYPU_2970	AMINO ACID PERMEASE	aapA	521	MCA_16
MYPU_3040	TRNA/RRNA METHYLTRANSFERASE	-	250	MCA_268
MYPU_3070	conserved hypothetical protein	-	289	MCA_33
MYPU_3210	OLIGOENDOPEPTIDASE F	pepF	613	MCA_508
MYPU_3260	conserved hypothetical protein	-	177	MCA_454
MYPU_3270	COMPETENCE-DAMAGE PROTEIN	cinA	133	MCA_196
MYPU_3460	ABC TRANSPORTER XYLOSE-BINDING LIPOPROTEIN	xylF	461	MCA_647
MYPU_3510	VACB-LIKE (Shigella flexneri) RIBONUCLEASE II	vacB	725	MCA_408
MYPU_3620	GLUCOSAMINE-6-PHOSPHATE ISOMERASE (GLUCOSAMINE-6-PHOSPHATE DEAMINASE) (GNPDA) (GLCN6P DEAMINASE)glucosamine	nagB	256	MCA_4
MYPU_3630	N-ACETYLMANNOSAMINE-6-P EPIMERASE	-	228	MCA_559
MYPU_3690	N-ACETYLGLUCOSAMINE-6-PHOSPHATE DEACETYLASE (GLCNAC 6-P DEACETYLASE)	nagA	253	MCA_553
MYPU_3700	unknown; predicted coding region	-	292	MCA_530
MYPU_3780	conserved hypothetical protein	-	149	MCA_229
MYPU_3830	TRSE-LIKE PROTEIN	trsE	853	MCA_467
MYPU_4140	predicted coding region	-	2244	MCA_616
MYPU_4150	LIPOPROTEIN	-	904	MCA_615
MYPU_4380	conserved hypothetical protein	-	792	MCA_478
MYPU_4390	unknown; predicted coding region	-	666	MCA_593
MYPU_4410	GLYCINE CLEAVAGE SYSTEM H PROTEIN	gcdh	111	MCA_550
MYPU_4420	unknown; predicted coding region	-	282	MCA_549
MYPU_4430	LIPOATE-PROTEIN LIGASE A	lplA	344	MCA_548
MYPU_4560	conserved hypothetical protein	-	251	MCA_236
MYPU_4570	unknown; predicted coding region	-	162	MCA_267
MYPU_4640	predicted coding region	-	450	MCA_531
MYPU_4750	conserved hypothetical protein	-	186	MCA_227
MYPU_4980	ABC TRANSPORTER PERMEASE PROTEIN	-	336	MCA_273
MYPU_5000	unknown; predicted coding region	-	736	MCA_568
MYPU_5010	unknown; predicted coding region	-	157	MCA_569
MYPU_5020	unknown; predicted coding region	-	326	MCA_570
MYPU_5030	unknown; predicted coding region	-	492	MCA_572
MYPU_5360	conserved hypothetical protein	-	435	MCA_628
MYPU_5390	RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE BETA CHAIN (RIBONUCLEOTIDE REDUCTASE)	nrdF	341	MCA_407
MYPU_5410	RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE ALPHA CHAIN (RIBONUCLEOTIDE REDUCTASE)	nrdE	711	MCA_405
MYPU_5520	DEOXYGUANOSINE KINASE (DGUO KINASE) (DGK) (DEOXYNUCLEOSIDE KINASE COMPLEX I F-COMPONENT)	dgk	216	MCA_181

Supplementary table E.3 continued

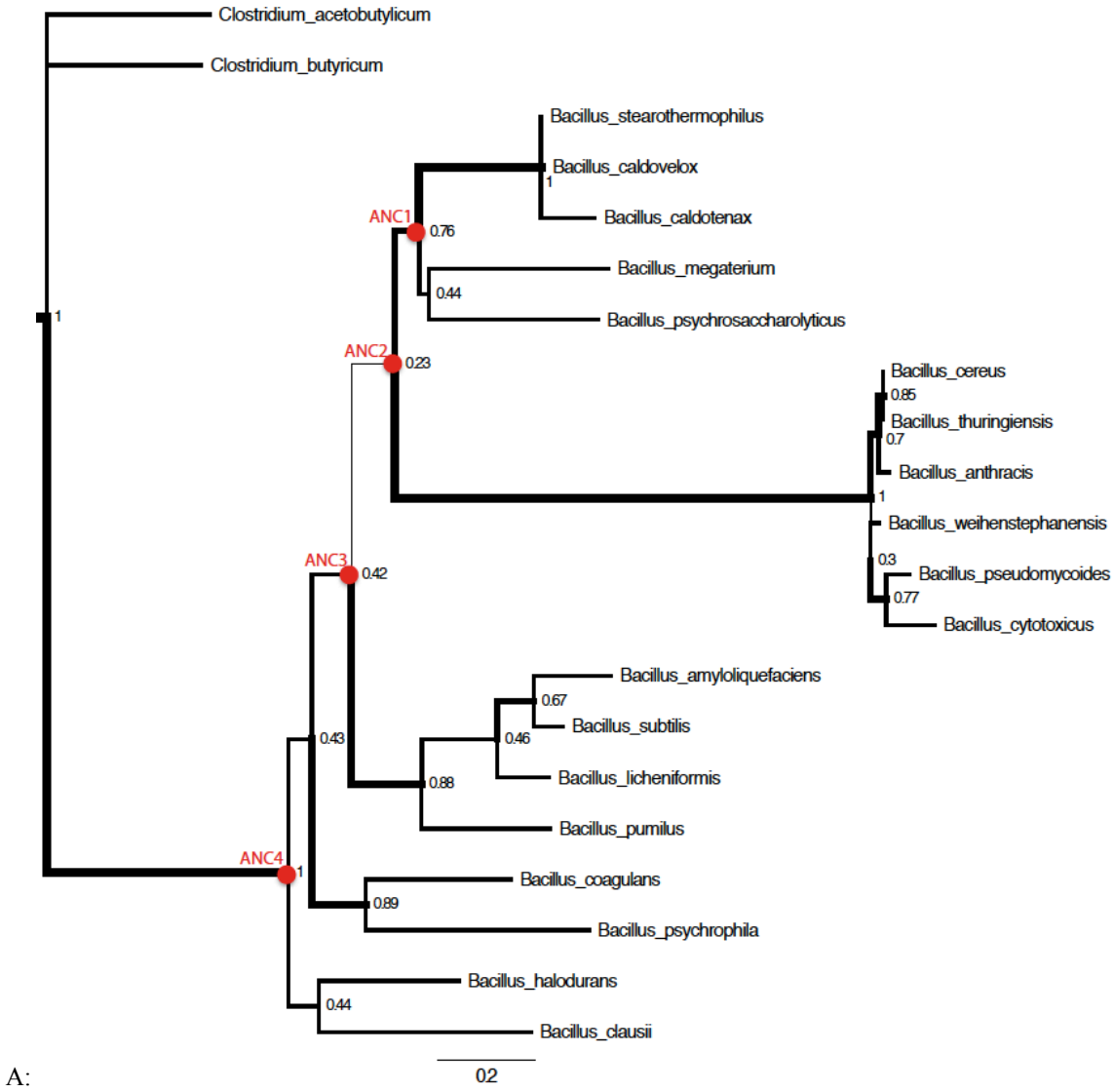
MYPU_5960	PENTITOL PHOSPHOTRANSFERASE ENZYME II, A COMPONENT	sgaA	159	MCA_536
MYPU_6090	30S RIBOSOMAL PROTEIN S18	rpsR	105	MCA_28
MYPU_6150	URACIL-DNA GLYCOSYLASE (UDG)	ung	221	MCA_257
MYPU_6170	SUGAR ABC TRANSPORTER PERMEASE PROTEIN	-	312	MCA_24
MYPU_6180	SUGAR ABC TRANSPORTER PERMEASE PROTEIN	-	598	MCA_25
MYPU_6220	Conserved hypothetical protein	-	315	MCA_652
MYPU_6240	conserved hypothetical protein	-	241	MCA_630
MYPU_6430	unknown; predicted coding region	-	189	MCA_629
MYPU_6570	HOLLIDAY JUNCTION DNA HELICASE RUVA	ruvA	197	MCA_278
MYPU_6600	predicted coding region	-	153	MCA_560
MYPU_6730	conserved hypothetical protein	-	79	MCA_253
MYPU_6830	RIBULOSE-PHOSPHATE 3-EPIMERASE (PENTOSE-5-PHOSPHATE 3-EPIMERASE) (PPE) (R5P3E)	rpe	215	MCA_120
MYPU_6840	conserved hypothetical protein	-	272	MCA_119
MYPU_6880	DNA METHYLASE	-	182	MCA_111
MYPU_6960	50S RIBOSOMAL PROTEIN L28	rpmB	63	MCA_247
MYPU_7010	unknown; predicted coding region	-	214	MCA_571
MYPU_7250	MANNOSE-6-PHOSPHATE ISOMERASE (PHOSPHOMANNOSE ISOMERASE) (PMI) (PHOSPHOHEXOMUTASE)	pmi	303	MCA_542
MYPU_7280	CHROMATE TRANSPORT PROTEIN	chrA	201	MCA_510
MYPU_7400	PTS SYSTEM, LICHENAN-SPECIFIC IIA COMPONENT	licA	266	MCA_627
MYPU_7480	HEXOSEPHOSPHATE TRANSPORT PROTEIN	uhpT	475	MCA_624
MYPU_7500	MANNITOL-1-PHOSPHATE 5-DEHYROGENASE	mtlD	360	MCA_482
MYPU_7510	PTS SYSTEM, MANNITOL-SPECIFIC IIABC COMPONENT (EIIABC-MTL) (MANNITOL-PERMEASE IIABC COMPONENT) (PHOSPHOTRANSFERASE ENZYME II, BC COMPONENT) (EII-MTL)	mtlA	342	MCA_485
MYPU_7560	TRANSPOSASE FOR INSERTION SEQUENCE ELEMENT IS1138 (Mycoplasma pulmonis)	-	402	MCA_667$
MYPU_7590	L-LACTATE DEHYDROGENASE	ldh	315	MCA_274
MYPU_7630	PYRUVATE DEHYDROGENASE E1 COMPONENT, BETA SUBUNIT	pdhB	332	MCA_141
MYPU_7720	NADH-DEPENDENT FLAVIN OXIDOREDUCTASE	baiH	398	MCA_547

Supplementary table E.4: List of 23 genes in the MCA ancestor, but absent in the hypothetical minimal genomes of either *M. pulmonis* or *M. genitalium*.

MG	product	gene symbol	length(aa)	MYPU	MCA
MG_009	deoxyribonuclease, TatD family, putative	-	262	MYPU_0140	MCA_441
MG_033	glycerol uptake facilitator	glpF	258	MYPU_2220	MCA_524
MG_039	FAD-dependent glycerol-3-phosphate dehydrogenase, putative	-	384	MYPU_2640	MCA_522
MG_056	tetrapyrrole (corrin/porphyrin) methylase protein	-	277	MYPU_0540	MCA_288
MG_061	Mycoplasma MFS transporter	-	567	MYPU_7480	MCA_624
MG_062	PTS system, fructose-specific IIABC component	fruA	680	MYPU_1290	MCA_399
MG_103	conserved hypothetical protein	-	280	MYPU_2260	MCA_452
MG_110	ribosome small subunit-dependent GTPase A	rsgA	278	MYPU_6840	MCA_119
MG_112	ribulose-phosphate 3-epimerase	rpe	209	MYPU_6830	MCA_120
MG_115	competence/damage-inducible protein CinA domain protein	-	157	MYPU_3270	MCA_196
MG_121	ABC transporter, permease protein	-	306	MYPU_6170	MCA_24
MG_183	oligoendopeptidase F	-	607	MYPU_3210	MCA_508
MG_252	RNA methyltransferase, TrmH family, group 3	-	242	MYPU_1780	MCA_434
MG_291	phosphonate ABC transporter, permease protein (P69), putative	-	543	MYPU_2180	MCA_597
MG_293	glycerophosphoryl diester phosphodiesterase family protein	-	244	MYPU_2630	MCA_506
MG_339	recA protein (recombinase A)	recA	340	MYPU_2520	MCA_197
MG_352	recombination protein U	recU	166	MYPU_0860	MCA_194
MG_360	ImpB/MucB/SamB family protein	-	411	MYPU_1880	MCA_281
MG_367	ribonuclease III	rnc	262	MYPU_1630	MCA_243
MG_380	methyltransferase GidB	-	192	MYPU_1280	MCA_456
MG_398	ATP synthase F1, epsilon subunit	atpC	133	MYPU_2650	MCA_413
MG_460	L-lactate dehydrogenase/malate dehydrogenase	ldh	312	MYPU_7590	MCA_274
MG_463	dimethyladenosine transferase	-	259	MYPU_0150	MCA_21

APPENDIX F

SUPPLEMENTARY INFORMATION FOR CHAPTER 7

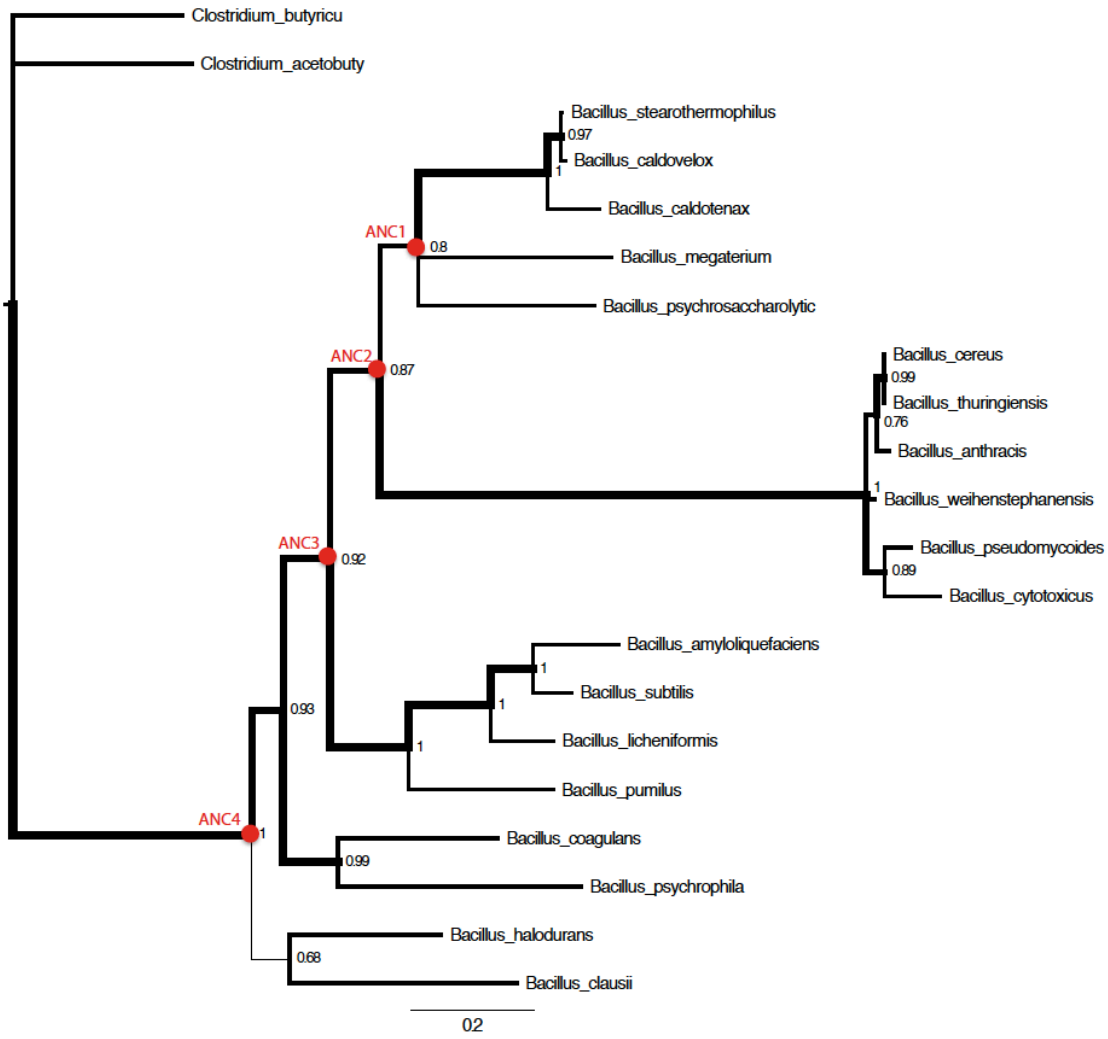


Supplementary figure F.1: Maximum likelihood and Bayesian phylogenies of LeuB proteins by Garli and MrBayes respectively.

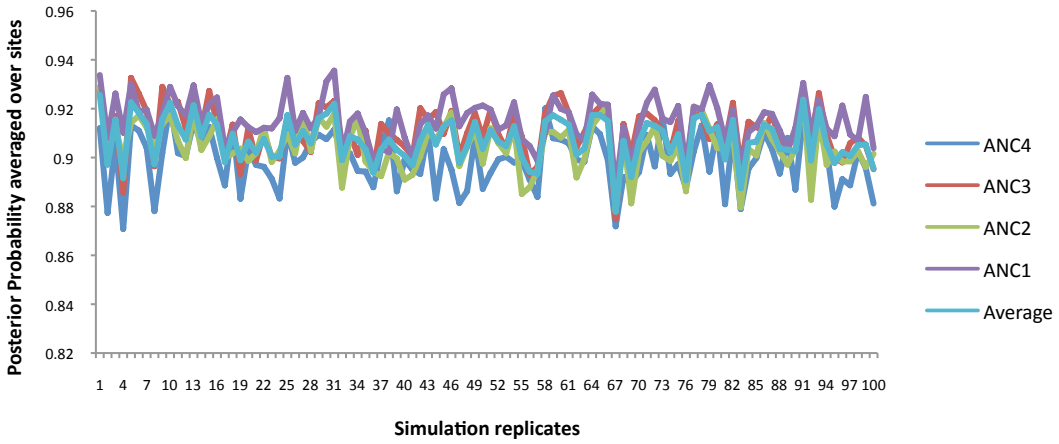
A: Maximum likelihood phylogeny of LeuB using Garli labeled with bootstrap values mapped upon the Hobbs' LeuB tree topology, and the width of the branches is based on bootstrap values.

B: Bayesian phylogeny of LeuB using MrBayes labeled with posterior probability values, and the width of the branches is based on posterior probability values.

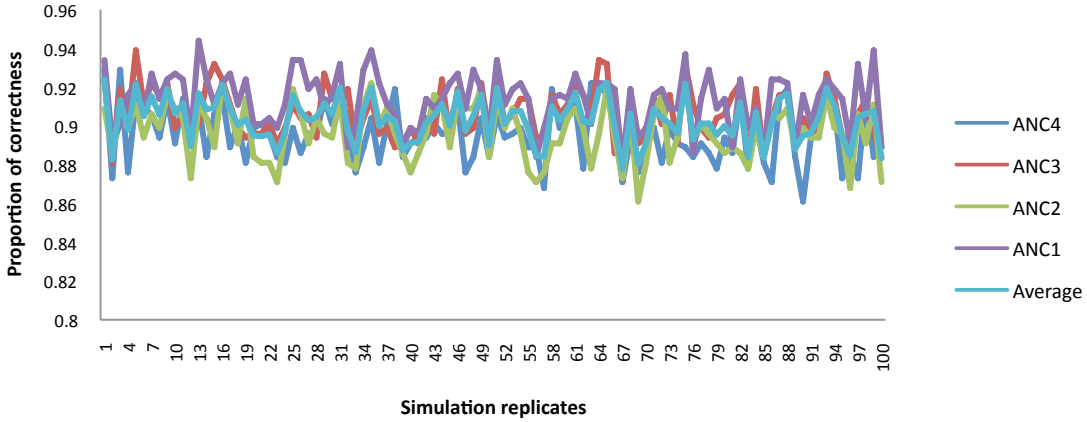
B:



Supplementary figure F.1 continued



A:

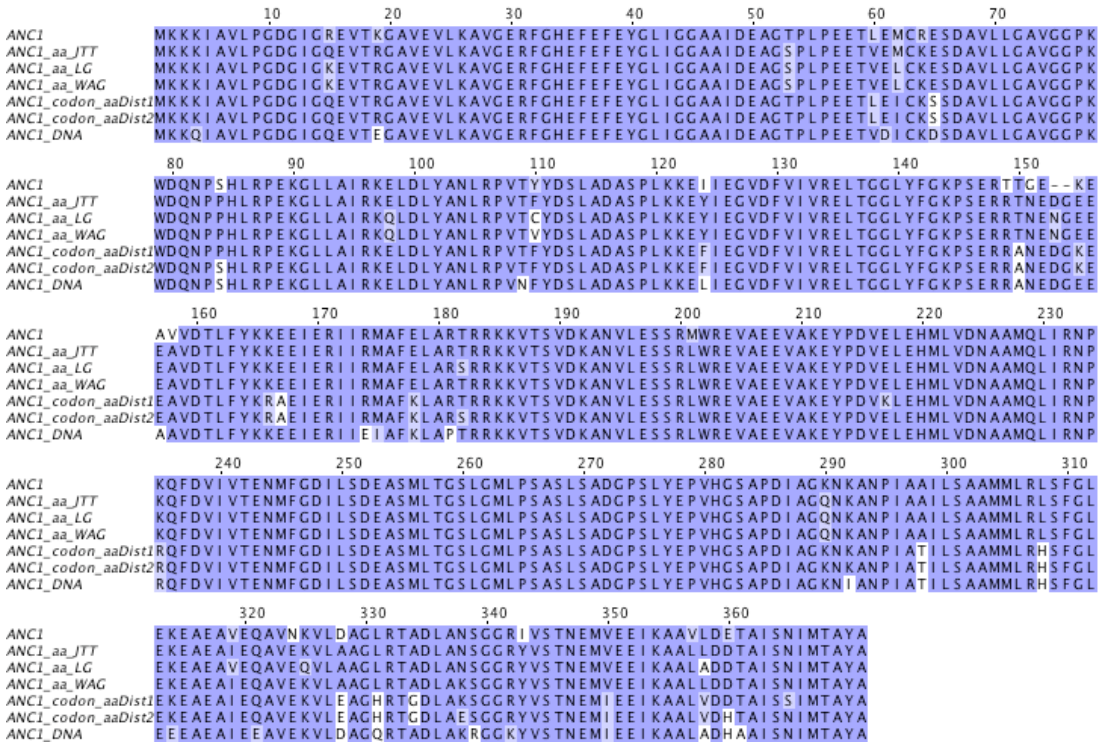


B:

Supplementary figure F.2: Accuracy and correctness of ancestral sequence reconstruction based on computational simulations.

A: The trend of accuracy of ancestral sequence reconstruction for 100 replicates using PAML for the four ancestral nodes from ANC1 to ANC4.

B: The trend of correctness for 100 replicates by comparing reconstructed ancestral sequences using PAML with the true ancestral sequences from the simulated data.



Supplementary figure F.3: Multiple sequence alignment of computationally reconstructed LeuB ancestral sequences by us using different models and datasets and by Hobbs.

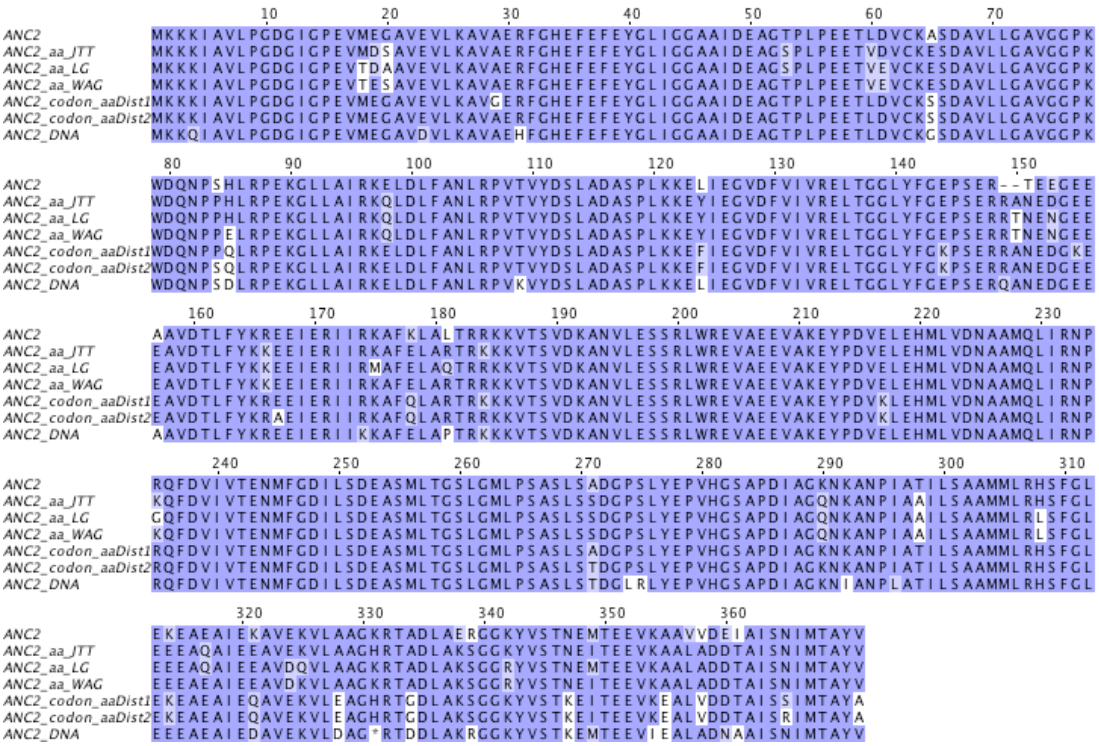
A: ANC1 multiple sequence alignment.

B: ANC2 multiple sequence alignment.

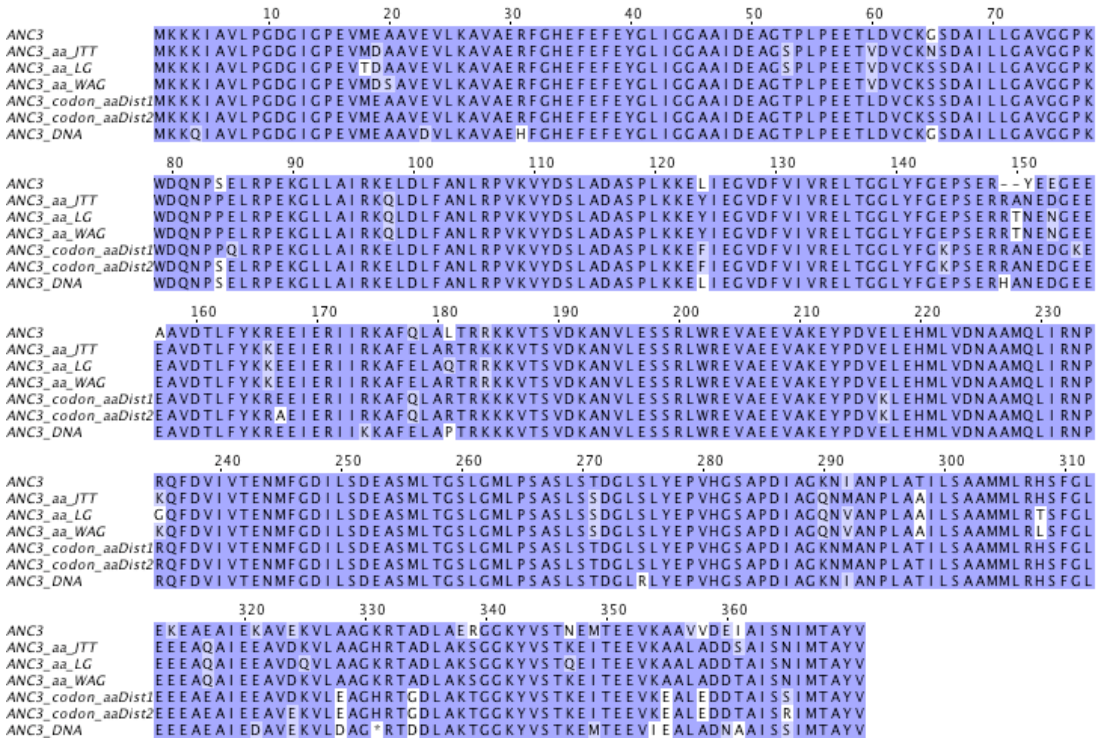
C: ANC3 multiple sequence alignment.

D: ANC4 multiple sequence alignment.

B:

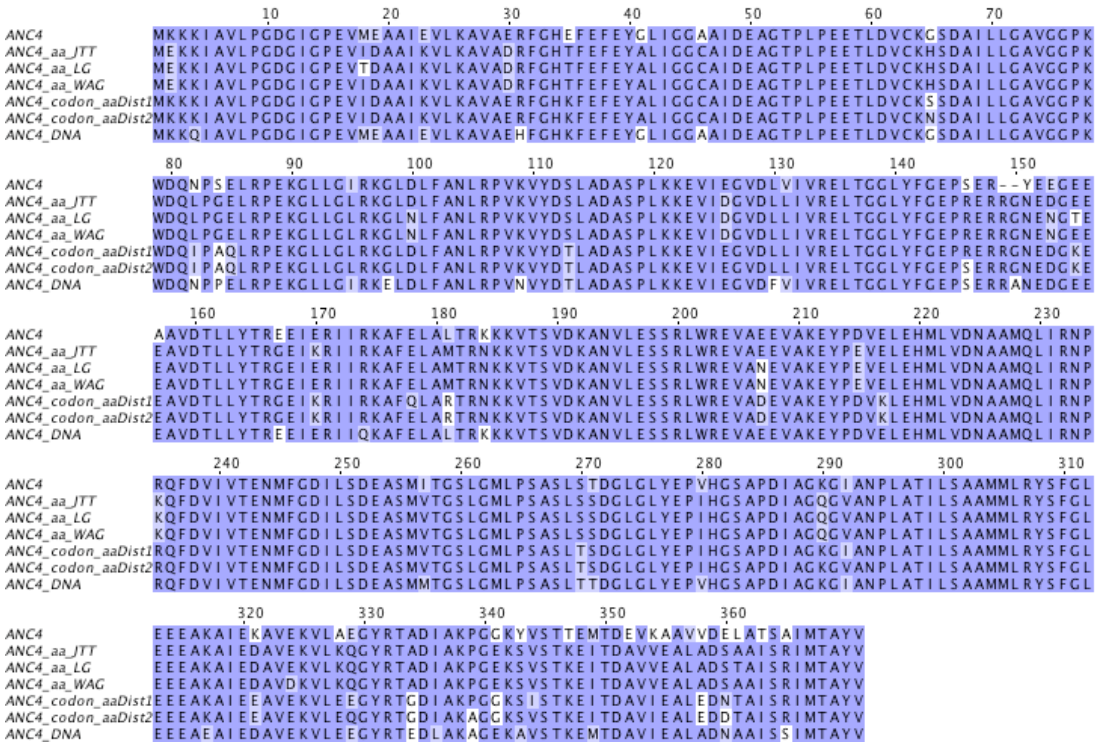


C:



Supplementary figure F.3 continued

D:



Supplementary figure F.3 continued

PUBLICATIONS

Z.-M. Zhao and E. A. Gaucher. ***Comparative genomics and genome evolution of Mycoplasmas via ancestral genome reconstruction***. (In Preparation)

Z.-M. Zhao and E. A. Gaucher. ***Computational validations of ancestral sequence reconstruction methods.*** (In Preparation)

Z.-M. Zhao and D. F. Mejia. ***The evolutionary history of bone morphogenetic proteins.*** (In Preparation)

D. F. Mejia, Z.-M. Zhao, and J. M. Sanchez-Ruiz. ***The ancestral resurrection of bone morphogenetic proteins.*** (In Preparation)

Z.-M. Zhao, K. Tang and XF Wu. ***Generating vaccines against Lagos bat viruses using ancestral sequence reconstruction.*** (In Preparation)

B. Kacar, Z.-M. Zhao, and etc. ***The origins of bacterial cytoskeleton organization***. (In Preparation)

R. Randall, Z.-M. Zhao, and E. A. Gaucher. ***Benchmark of computational ancestral sequence reconstruction methods***. (In Preparation)

Z.-M. Zhao, A. B. Reynolds, and E. A. Gaucher. ***The evolutionary history of the catenin gene family during metazoan evolution.*** BMC Evol. Biol. 2011. (* Highly accessed)

R. Perez-Jimenez, A. Inglés-Prieto, Z.-M. Zhao, I. Sanchez-Romero, J. Alegre-Cebollada, P. Kosuri, S. Garcia-Manyes, J. M. Sanchez-Ruiz, E. A. Gaucher; J. M. Fernandez. ***Paleoenzymology at the single-molecule level: probing the chemistry of resurrected enzymes.*** Nature Struct. & Mol. Biol. 2011.

X.-F. Wan, M. Emch, and Z.-M. Zhao. ***Advances in molecular evolution of influenza A viruses***. In Global View at Fight against Influenza, edited by Mitrasinovic. 2009. (Book Chapter)

X.-F. Wan, T. Nguyen, T. C. Davis, C. B. Smith, Z.-M. Zhao, M. Carrel, S. Jadhao, A. Balish, F. Luo, M. Emch, Y. Matsuoka, N. J. Cox, A. Klimov, and R. O. Donis. ***Evolution of Highly Pathogenic H5N1 Avian Influenza Viruses in Vietnam between 2001 and 2007***. PLoS One. 2008.

Z.-M. Zhao, K. F. Shortridge, M. Garcia, Y. Guan and X.-F. Wan. ***Genotypic diversity of H5N1 highly pathogenic avian influenza viruses.*** J Gen Virol. 2008.

REFERENCES

1. Graur D, Li W-H: **Fundamentals of Molecular Evolution**: Sinauer Associates; 2000.

2. Posada D: **jModelTest: phylogenetic model averaging**. *Mol Biol Evol* 2008, **25**(7):1253-1256.

3. Nylander JAA: **MrModeltest v2. Program distributed by the author.** . In.: Evolutionary Biology Centre, Uppsala University. ; 2004.

4. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution**. *Bioinformatics* 2005, **21**(9):2104-2105.

5. Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution**. *Bioinformatics*, **27**(8):1164-1165.

6. Yang Z, Rannala B: **Molecular phylogenetics: principles and practice**. *Nat Rev Genet*, **13**(5):303-314.

7. Hillis DM, Bull JJ: **An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis**. *Systematic Biology* 1993, **42**(2):182-192.

8. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms**. *Proc Natl Acad Sci U S A* 1977, **74**(11):5088-5090.

9. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML: **Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences**. *Proceedings of the National Academy of Sciences* 2010.

10. Kumar S, Dudley JT, Filipski A, Liu L: **Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations**. *Trends Genet*, **27**(9):377-386.

11. Zhang J: **Evolution by gene duplication: an update**. *Trends in Ecology & Evolution* 2003, **18**(6):292-298.

12. Koonin EV: **Orthologs, paralogs, and evolutionary genomics**. *Annual review of genetics* 2005, **39**:309-338.

13. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life**. *Science* 2006, **311**(5765):1283-1287.

14. Snel B, Bork P, Huynen MA: **Genome phylogeny based on gene content**. *Nature genetics* 1999, **21**(1):108-110.

15. Sankoff D, Blanchette M: **Multiple genome rearrangement and breakpoint phylogeny**. *Journal of computational biology : a journal of computational molecular cell biology* 1998, **5**(3):555-570.

16. Tekaia F, Yeramian E: **Genome trees from conservation profiles**. *PLoS Comput Biol* 2005, **1**(7):e75.

17. Zhao ZM, Reynolds AB, Gaucher EA: **The evolutionary history of the catenin gene family during metazoan evolution**. *BMC Evol Biol*, **11**:198.

18. Brown JR, Doolittle WF: **Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications**. *Proc Natl Acad Sci U S A* 1995, **92**(7):2441-2445.

19. Hurles M: **Gene duplication: the genomic trade in spare parts**. *PLoS Biol* 2004, **2**(7):E206.

20. Gu X: **A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences**. *Mol Biol Evol* 2006, **23**(10):1937-1945.

21. Holmes EC, Ghedin E, Miller N, Taylor J, Bao Y, St George K, Grenfell BT, Salzberg SL, Fraser CM, Lipman DJ, Taubenberger JK: **Whole-Genome Analysis of Human Influenza A Virus Reveals Multiple Persistent Lineages and Reassortment among Recent H3N2 Viruses**. *PLoS Biol* 2005, **3**(9):e300.

22. **Influenza Virology: Current Topics**: Caister Academic Press; 2006.

23. **Global View of the Fight Against Influenza**: Nova Science Pub Inc 2009.

24. Hay AJ, Gregory V, Douglas AR, Lin YP: **The evolution of human influenza viruses**. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2001, **356**(1416):1861-1870.

25. Dugan VG, Chen R, Spiro DJ, Sengamalay N, Zaborsky J, Ghedin E, Nolting J, Swayne DE, Runstadler JA, Happ GM, Senne DA, Wang R, Slemons RD, Holmes EC, Taubenberger JK: **The evolutionary genetics and emergence of avian influenza viruses in wild birds**. *PLoS pathogens* 2008, **4**(5):e1000076.

26. Guan Y, Peiris JS, Lipatov AS, Ellis TM, Dyrting KC, Krauss S, Zhang LJ, Webster RG, Shortridge KF: **Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(13):8950-8955.

27. Banks J, Speidel ES, Moore E, Plowright L, Piccirillo A, Capua I, Cordioli P, Fioretti A, Alexander DJ: **Changes in the haemagglutinin and the neuraminidase genes prior to the emergence of highly pathogenic H7N1 avian influenza viruses in Italy**. *Archives of virology* 2001, **146**(5):963-973.

28. Tarendeau F, Crepin T, Guilligay D, Ruigrok RW, Cusack S, Hart DJ: **Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit**. *PLoS pathogens* 2008, **4**(8):e1000136.

29. Hungnes O: **The role of genetic analysis in influenza virus surveillance and strain characterisation**. *Vaccine* 2002, **20 Suppl 5**:B45-49.

30. Galsworthy SJ, ten Bosch QA, Hoye BJ, Heesterbeek JA, Klaassen M, Klinkenberg D: **Effects of infection-induced migration delays on the epidemiology of avian influenza in wild mallard populations**. *PLoS One* 2011, **6**(10):e26118.

31. Zhao ZM, Shortridge KF, Garcia M, Guan Y, Wan XF: **Genotypic diversity of H5N1 highly pathogenic avian influenza viruses**. *J Gen Virol* 2008, **89**(Pt 9):2182-2193.

32. Pauling L, Zuckerkandl E: **Chemical paleogenetics: molecular "restoration studies" of extinct forms of life**. *Acta Chemica Scandinavica* 1963, **17 supl.**

33. Thornton JW: **Resurrecting ancient genes: experimental analysis of extinct molecules**. *Nat Rev Genet* 2004, **5**(5):366-375.

34. Thornton JW, Need E, Crews D: **Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling**. *Science (New York, NY* 2003, **301**(5640):1714-1717.

35. Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA: **Resurrecting ancestral alcohol dehydrogenases from yeast**. *Nature genetics* 2005, **37**(6):630-635.

36. Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M: **Parallel adaptations to high temperatures in the Archaean eon**. *Nature* 2008, **456**(7224):942-945.

37. Chang BS, Jonsson K, Kazmi MA, Donoghue MJ, Sakmar TP: **Recreating a functional ancestral archosaur visual pigment**. *Molecular biology and evolution* 2002, **19**(9):1483-1489.

38. Perez-Jimenez R, Ingles-Prieto A, Zhao ZM, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A, Sanchez-Ruiz JM, Gaucher EA, Fernandez JM: **Single-molecule paleoenzymology probes the chemistry of resurrected enzymes**. *Nat Struct Mol Biol*, **18**(5):592-596.

39. Gaucher EA, Thomson JM, Burgan MF, Benner SA: **Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins**. *Nature* 2003, **425**(6955):285-288.

40. Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA: **The ribonuclease from an extinct bovid ruminant**. *FEBS Lett* 1990, **262**(1):104-106.

41. Hillis DM, Bull JJ, White ME, Badgett MR, Molineux IJ: **Experimental phylogenetics: generation of a known phylogeny**. *Science* 1992, **255**(5044):589-592.

42. Yang Z, Kumar S, Nei M: **A new method of inference of ancestral nucleotide and amino acid sequences**. *Genetics* 1995, **141**(4):1641-1650.

43. Williams PD, Pollock DD, Blackburne BP, Goldstein RA: **Assessing the accuracy of ancestral protein reconstruction methods**. *PLoS Comput Biol* 2006, **2**(6):e69.

44. Hanson-Smith V, Kolaczkowski B, Thornton JW: **Robustness of ancestral sequence reconstruction to phylogenetic uncertainty**. *Mol Biol Evol*, **27**(9):1988-1999.

45. Wan X-F: **Isolation and characterization of avian influenza viruses in China.** . *Master thesis*. Guangzhou: South China Agricultural University; 1998.

46. Guo Y, Xu X, Wan X: **[Genetic characterization of an avian influenza A (H5N1) virus isolated from a sick goose in China]**. *Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi* 1998, **12**(4):322-325.

47. Claas EC, Osterhaus AD, van Beek R, De Jong JC, Rimmelzwaan GF, Senne DA, Krauss S, Shortridge KF, Webster RG: **Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus**. *Lancet* 1998, **351**(9101):472-477.

48. Sims LD, Ellis TM, Liu KK, Dyrting K, Wong H, Peiris M, Guan Y, Shortridge KF: **Avian influenza in Hong Kong 1997-2002**. *Avian diseases* 2003, **47**(3 Suppl):832-838.

49. Shortridge KF, Zhou NN, Guan Y, Gao P, Ito T, Kawaoka Y, Kodihalli S, Krauss S, Markwell D, Murti KG, Norwood M, Senne D, Sims L, Takada A, Webster RG: **Characterization of avian H5N1 influenza viruses from poultry in Hong Kong**. *Virology* 1998, **252**(2):331-342.

50. Chen H, Smith GJ, Zhang SY, Qin K, Wang J, Li KS, Webster RG, Peiris JS, Guan Y: **Avian flu: H5N1 virus outbreak in migratory waterfowl**. *Nature* 2005, **436**(7048):191-192.

51. Liu J, Xiao H, Lei F, Zhu Q, Qin K, Zhang XW, Zhang XL, Zhao D, Wang G, Feng Y, Ma J, Liu W, Wang J, Gao GF: **Highly pathogenic H5N1 influenza virus infection in migratory birds**. *Science (New York, NY* 2005, **309**(5738):1206.

52. Shortridge KF: **Pandemic influenza: a zoonosis?** *Seminars in respiratory infections* 1992, **7**(1):11-25.

53. Shortridge KF, Stuart-Harris CH: **An influenza epicentre?** *Lancet* 1982, **2**(8302):812-813.

54. Xu X, Subbarao, Cox NJ, Guo Y: **Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong**. *Virology* 1999, **261**(1):15-19.

55. Guan Y, Poon LL, Cheung CY, Ellis TM, Lim W, Lipatov AS, Chan KH, Sturm-Ramirez KM, Cheung CL, Leung YH, Yuen KY, Webster RG, Peiris JS: **H5N1 influenza: a protean pandemic threat**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(21):8156-8161.

56. Chen H, Deng G, Li Z, Tian G, Li Y, Jiao P, Zhang L, Liu Z, Webster RG, Yu K: **The evolution of H5N1 influenza viruses in ducks in southern China**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(28):10452-10457.

57. Chen H, Smith GJ, Li KS, Wang J, Fan XH, Rayner JM, Vijaykrishna D, Zhang JX, Zhang LJ, Guo CT, Cheung CL, Xu KM, Duan L, Huang K, Qin K, Leung YH, Wu WL, Lu HR, Chen Y, Xia NS, Naipospos TS, Yuen KY, Hassan SS, Bahri S, Nguyen TD,

Webster RG, Peiris JS, Guan Y: **Establishment of multiple sublineages of H5N1 influenza virus in Asia: Implications for pandemic control**. *Proceedings of the National Academy of Sciences of the United States of America* 2006.

58. Li KS, Guan Y, Wang J, Smith GJ, Xu KM, Duan L, Rahardjo AP, Puthavathana P, Buranathai C, Nguyen TD, Estoepangestie AT, Chaisingh A, Auewarakul P, Long HT, Hanh NT, Webby RJ, Poon LL, Chen H, Shortridge KF, Yuen KY, Webster RG, Peiris JS: **Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia**. *Nature* 2004, **430**(6996):209-213.

59. Wan XF, Chen G, Luo F, Emch M, Donis R: **A quantitative genotype algorithm reflecting H5N1 Avian influenza niches**. *Bioinformatics* 2007, **23**(18):2368-2375.

60. Duan L, Campitelli L, Fan XH, Leung YH, Vijaykrishna D, Zhang JX, Donatelli I, Delogu M, Li KS, Foni E, Chiapponi C, Wu WL, Kai H, Webster RG, Shortridge KF, Peiris JS, Smith GJ, Chen H, Guan Y: **Characterization of low-pathogenic H5 subtype influenza viruses from Eurasia: implications for the origin of highly pathogenic H5N1 viruses**. *Journal of virology* 2007, **81**(14):7529-7539.

61. Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T: **National center for biotechnology information viral genomes project**. *Journal of virology* 2004, **78**(14):7291-7298.

62. Wan X-F, Wu X, Lin G, Holton SB, Desmone RA, Shyu CR, Guan Y, Emch M: **Computational Identification of Reassortments in Avian Influenza Viruses**. *Avian Diseases* 2007, **51**:434-439.

63. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

64. Swofford DL: **PAUP*: Phylogenic analysis using Parsimony**: Sinauer,. Sunderland, Massachusetts; 1998.

65. Zwickl DJ: **Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion**. The University of Texas at Austin.; 2006.

66. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees**. *BMC Evol Biol* 2007, **7**:214.

67. Posada D, Crandall KA: **MODELTEST: testing the model of DNA substitution**. *Bioinformatics* 1998, **14**(9):817-818.

68. Mukhtar MM, Rasool ST, Song D, Zhu C, Hao Q, Zhu Y, Wu J: **Origin of highly pathogenic H5N1 avian influenza virus in China and genetic characterization of donor and recipient viruses**. *The Journal of general virology* 2007, **88**(Pt 11):3094-3099.

69. Smith GJ, Fan XH, Wang J, Li KS, Qin K, Zhang JX, Vijaykrishna D, Cheung CL, Huang K, Rayner JM, Peiris JS, Chen H, Webster RG, Guan Y: **Emergence and**

predominance of an H5N1 influenza variant in China. *Proc Natl Acad Sci U S A* 2006, **103**(45):16936-16941.

70. Okazaki K, Takada A, Ito T, Imai M, Takakuwa H, Hatta M, Ozaki H, Tanizaki T, Nagano T, Ninomiya A, Demenev VA, Tyaptirganov MM, Karatayeva TD, Yamnikova SS, Lvov DK, Kida H: **Precursor genes of future pandemic influenza viruses are perpetuated in ducks nesting in Siberia**. *Archives of virology* 2000, **145**(5):885-893.

71. Twu KY, Kuo RL, Marklund J, Krug RM: **The H5N1 influenza virus NS genes selected after 1998 enhance virus replication in mammalian cells**. *Journal of virology* 2007, **81**(15):8112-8121.

72. Seo SH, Hoffmann E, Webster RG: **Lethal H5N1 influenza viruses escape host anti-viral cytokine responses**. *Nature medicine* 2002, **8**(9):950-954.

73. Wan XF, Ren T, Luo KJ, Liao M, Zhang GH, Chen JD, Cao WS, Li Y, Jin NY, Xu D, Xin CA: **Genetic characterization of H5N1 avian influenza viruses isolated in southern China during the 2003-04 avian influenza outbreaks**. *Archives of virology* 2005, **150**(6):1257-1266.

74. Long JX, Xue F, Peng Y, Gu M, Liu XF: **[The deletion of nucleotides of NS gene from 263 to 277 of H5N1 increases viral virulence in chicken]**. *Wei Sheng Wu Xue Bao* 2006, **46**(2):301-305.

75. Li Y, Lin Z, Shi J, Qi Q, Deng G, Li Z, Wang X, Tian G, Chen H: **Detection of Hong Kong 97-like H5N1 influenza viruses from eggs of Vietnamese waterfowl**. *Arch Virol* 2006, **151**(8):1615-1624.

76. Wan XF, Nguyen T, Davis CT, Smith CB, Zhao ZM, Carrel M, Inui K, Do HT, Mai DT, Jadhao S, Balish A, Shu B, Luo F, Emch M, Matsuoka Y, Lindstrom SE, Cox NJ, Nguyen CV, Klimov A, Donis RO: **Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2007**. *PLoS One* 2008, **3**(10):e3462.

77. Nguyen DC, Uyeki TM, Jadhao S, Maines T, Shaw M, Matsuoka Y, Smith C, Rowe T, Lu X, Hall H, Xu X, Balish A, Klimov A, Tumpey TM, Swayne DE, Huynh LP, Nghiem HK, Nguyen HH, Hoang LT, Cox NJ, Katz JM: **Isolation and characterization of avian influenza viruses, including highly pathogenic H5N1, from poultry in live bird markets in Hanoi, Vietnam, in 2001**. *J Virol* 2005, **79**(7):4201-4212.

78. Wan X-F, Ozden M, Lin G: **Ubiquitous reassortments in influenza A viruses.** *J Bioinform Comput Biol* 2008, **In press.**

79. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics* 2003, **19**(12):1572-1574.

80. WHOOIEFAOH5N1EvolutionWorkingGroup: **Towards a unified nomenclature system for the highly pathogenic avian influenza H5N1 viruses.** *Emerg Infect Dis* 2008, **In press**.

81. McCrea PD, Gu D: **The catenin family at a glance**. *J Cell Sci*, **123**(Pt 5):637-642.

82. Anastasiadis PZ, Reynolds AB: **The p120 catenin family: complex roles in adhesion, signaling and cancer**. *J Cell Sci* 2000, **113 (Pt 8)**:1319-1334.

83. Schneider SQ, Finnerty JR, Martindale MQ: **Protein evolution: structure-function relationships of the oncogene beta-catenin in the evolution of multicellular animals**. *J Exp Zool B Mol Dev Evol* 2003, **295**(1):25-44.

84. Kobielak A, Fuchs E: **Alpha-catenin: at the junction of intercellular adhesion and actin dynamics**. *Nat Rev Mol Cell Biol* 2004, **5**(8):614-625.

85. Desai BV, Harmon RM, Green KJ: **Desmosomes at a glance**. *J Cell Sci* 2009, **122**(Pt 24):4401-4407.

86. McCrea PD, Park JI: **Developmental functions of the P120-catenin sub-family**. *Biochim Biophys Acta* 2007, **1773**(1):17-33.

87. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, Li W, Lyons JB, Morris A, Nichols S, Richter DJ, Salamov A, Sequencing JG, Bork P, Lim WA, Manning G *et al*: **The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans**. *Nature* 2008, **451**(7180):783-788.

88. Bocquet J, Winzenberg T, Shaw KA: **Epicentre of influenza - the primary care experience in Melbourne, Victoria**. *Aust Fam Physician*, **39**(5):313-316.

89. Darwin C: **On the origin of species**. United Kingdom: John Murray; 1859.

90. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948.

91. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F: **Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference**. *Bioinformatics* 2004, **20**(3):407-415.

92. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

93. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Mol Biol Evol* 2007, **24**(8):1596-1599.

94. Carrel MA, Emch M, Jobe RT, Moody A, Wan XF: **Spatiotemporal structure of molecular evolution of H5N1 highly pathogenic avian influenza viruses in Vietnam**. *PLoS One*, **5**(1):e8631.

95. Gu X, Vander Velden K: **DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family**. *Bioinformatics* 2002, **18**(3):500-501.

96. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW, 3rd, Su AI: **BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources**. *Genome Biol* 2009, **10**(11):R130.

97. Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R: **The new animal phylogeny: reliability and implications**. *Proc Natl Acad Sci U S A* 2000, **97**(9):4453-4456.

98. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Comput Appl Biosci* 1997, **13**(5):555-556.

99. Choi HJ, Weis WI: **Structure of the armadillo repeat domain of plakophilin 1**. *J Mol Biol* 2005, **346**(1):367-376.

100. Hobmayer E, Hatta M, Fischer R, Fujisawa T, Holstein TW, Sugiyama T: **Identification of a Hydra homologue of the beta-catenin/plakoglobin/armadillo gene family**. *Gene* 1996, **172**(1):155-159.

101. Coates JC, Harwood AJ: **Cell-cell adhesion and signal transduction during Dictyostelium development**. *J Cell Sci* 2001, **114**(Pt 24):4349-4358.

102. Tang F, Peng Y, Nau JJ, Kauffman EJ, Weisman LS: **Vac8p, an armadillo repeat protein, coordinates vacuole inheritance with multiple vacuolar processes**. *Traffic* 2006, **7**(10):1368-1377.

103. Mendel JG: **Experiments in plant hybridization**. 1865.

104. Mariner DJ, Wang J, Reynolds AB: **ARVCF localizes to the nucleus and adherens junction and is mutually exclusive with p120(ctn) in E-cadherin complexes**. *J Cell Sci* 2000, **113 (Pt 8)**:1481-1490.

105. Daniel JM: **Dancing in and out of the nucleus: p120(ctn) and the transcription factor Kaiso**. *Biochim Biophys Acta* 2007, **1773**(1):59-68.

106. Hosking CR, Ulloa F, Hogan C, Ferber EC, Figueroa A, Gevaert K, Birchmeier W, Briscoe J, Fujita Y: **The transcriptional repressor Glis2 is a novel binding partner for p120 catenin**. *Mol Biol Cell* 2007, **18**(5):1918-1927.

107. Fang X, Ji H, Kim SW, Park JI, Vaught TG, Anastasiadis PZ, Ciesiolka M, McCrea PD: **Vertebrate development requires ARVCF and p120 catenins and their interplay with RhoA and Rac**. *J Cell Biol* 2004, **165**(1):87-98.

108. Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW: **Evolution of increased complexity in a molecular machine**. *Nature*, **481**(7381):360-364.

109. Bridgham JT, Carroll SM, Thornton JW: **Evolution of hormone-receptor complexity by molecular exploitation**. *Science* 2006, **312**(5770):97-101.

110. Reynolds AB, Roczniak-Ferguson A: **Emerging roles for p120-catenin in cell adhesion and cancer**. *Oncogene* 2004, **23**(48):7947-7956.

111. Hatzfeld M: **Plakophilins: Multifunctional proteins or just regulators of desmosomal adhesion?** *Biochim Biophys Acta* 2007, **1773**(1):69-77.

112. S OO, Saitou N: **Phylogenetic relationship of muscle tissues deduced from superimposition of gene trees**. *Mol Biol Evol* 1999, **16**(6):856-867.

113. Eckhart L, Valle LD, Jaeger K, Ballaun C, Szabo S, Nardi A, Buchberger M, Hermann M, Alibardi L, Tschachler E: **Identification of reptilian genes encoding hair keratin-like proteins suggests a new scenario for the evolutionary origin of hair**. *Proc Natl Acad Sci U S A* 2008, **105**(47):18419-18423.

114. Shibuya Y, Murata M, Munemoto S, Masago H, Takeuchi J, Wada K, Yokoo S, Umeda M, Komori T: **Alpha E- and alpha N-catenin expression in dorsal root ganglia and spinal cord**. *Kobe J Med Sci* 2003, **49**(3-4):93-98.

115. Stocker AM, Chenn A: **Differential expression of alpha-E-catenin and alpha-N-catenin in the developing cerebral cortex**. *Brain Res* 2006, **1073-1074**:151-158.

116. Bogaerts S, Vanlandschoot A, van Hengel J, van Roy F: **Nuclear translocation of alphaN-catenin by the novel zinc finger transcriptional repressor ZASC1**. *Exp Cell Res* 2005, **311**(1):1-13.

117. Janssens B, Mohapatra B, Vatta M, Goossens S, Vanpoucke G, Kools P, Montoye T, van Hengel J, Bowles NE, van Roy F, Towbin JA: **Assessment of the CTNNA3 gene encoding human alpha T-catenin regarding its involvement in dilated cardiomyopathy**. *Hum Genet* 2003, **112**(3):227-236.

118. Goossens S, Janssens B, Bonne S, De Rycke R, Braet F, van Hengel J, van Roy F: **A unique and specific interaction between alphaT-catenin and plakophilin-2 in the area composita, the mixed-type junctional structure of cardiac intercalated discs**. *J Cell Sci* 2007, **120**(Pt 12):2126-2136.

119. Janssens B, Goossens S, Staes K, Gilbert B, van Hengel J, Colpaert C, Bruyneel E, Mareel M, van Roy F: **alphaT-catenin: a novel tissue-specific beta-catenin-binding protein mediating strong cell-cell adhesion**. *J Cell Sci* 2001, **114**(Pt 17):3177-3188.

120. Goossens S, Janssens B, Vanpoucke G, De Rycke R, van Hengel J, van Roy F: **Truncated isoform of mouse alphaT-catenin is testis-restricted in expression and function**. *FASEB J* 2007, **21**(3):647-655.

121. Perez-Jimenez R, Ingl √ ©s-Prieto A, Zhao Z-M, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A, Sanchez-Ruiz JM, Gaucher EA, Fernandez JM: **Single-molecule paleoenzymology probes the chemistry of resurrected enzymes**. *Nat Struct Mol Biol*, **18**(5):592-596.

122. Benner SA, Sassi SO, Gaucher EA: **Molecular paleoscience: systems biology from the past**. *Adv Enzymol Relat Areas Mol Biol* 2007, **75**:1-132, xi.

123. Nisbet EG, Sleep NH: **The habitat and nature of early life**. *Nature* 2001, **409**(6823):1083-1091.

124. Pollock DD, Chang BSW: **in Ancestral sequence reconstruction**. Oxford ; New York: ed. Liberles. D.A., Oxford University Press; 2007.

125. Gaucher EA, Govindarajan S, Ganesh OK: **Palaeotemperature trend for Precambrian life inferred from resurrected proteins**. *Nature* 2008, **451**(7179):704-707.

126. Holmgren A: **Thioredoxin**. *Annual review of biochemistry* 1985, **54**:237-271.

127. Swofford DL: **PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)** *Sunderland, MA: Sinauer Associates* 1998.

128. Battistuzzi FU, Feijao A, Hedges SB: **A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land**. *BMC evolutionary biology* 2004, **4**:44.

129. Raymond J, Segre D: **The effect of oxygen on biochemical networks and the evolution of complex life**. *Science (New York, NY* 2006, **311**(5768):1764-1767.

130. Liberles DA: **Ancestral sequence reconstruction**. Oxford ; New York: Oxford University Press; 2007.

131. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput Appl Biosci* 1992, **8**(3):275-282.

132. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA, Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA, 3rd, Smith HO, Venter JC: **Creation of a bacterial cell controlled by a chemically synthesized genome**. *Science*, **329**(5987):52-56.

133. Lartigue C, Glass JI, Alperovich N, Pieper R, Parmar PP, Hutchison CA, Smith HO, Venter JC: **Genome Transplantation in Bacteria: Changing One Species to Another**. *Science* 2007, **317**(5838):632-638.

134. Lartigue C, Vashee S, Algire MA, Chuang R-Y, Benders GA, Ma L, Noskov VN, Denisova EA, Gibson DG, Assad-Garcia N, Alperovich N, Thomas DW, Merryman C, Hutchison CA, Smith HO, Venter JC, Glass JI: **Creating Bacterial Strains from Genomes That Have Been Cloned and Engineered in Yeast**. *Science* 2009, **325**(5948):1693-1696.

135. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA, Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA, 3rd, Smith HO, Venter JC: **Creation of a bacterial cell controlled by a chemically synthesized genome**. *Science* 2010, **329**(5987):52-56.

136. Lee IM, Davis RE, Gundersen-Rindal DE: **Phytoplasma: phytopathogenic mollicutes**. *Annu Rev Microbiol* 2000, **54**:221-255.

137. Gasparich GE, Whitcomb RF, Dodge D, French FE, Glass J, Williamson DL: **The genus Spiroplasma and its non-helical descendants: phylogenetic classification, correlation with phenotype and roots of the Mycoplasma mycoides clade**. *Int J Syst Evol Microbiol* 2004, **54**(Pt 3):893-918.

138. Wolf M, Muller T, Dandekar T, Pollack JD: **Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data**. *Int J Syst Evol Microbiol* 2004, **54**(3):871-875.

139. Markov AV, Zakharov IA: **Evolution of gene orders in mycoplasmas (Bacteria, Firmicutes, Mollicutes)**. *Genetika* 2009, **45**(7):893-899.

140. Razin SH, Richard **Molecular biology and pathogenicity of mycoplasmas**: Springer; 2002.

141. Razin S, Yogev D, Naot Y: **Molecular biology and pathogenicity of mycoplasmas**. *Microbiol Mol Biol Rev* 1998, **62**(4):1094-1156.

142. Sirand-Pugnet P, Citti C, Barre A, Blanchard A: **Evolution of mollicutes: down a bumpy road with twists and turns**. *Res Microbiol* 2007, **158**(10):754-766.

143. Gasparich GE, Whitcomb RF, Dodge D, French FE, Glass J, Williamson DL: **The genus Spiroplasma and its non-helical descendants: phylogenetic classification, correlation with phenotype and roots of the Mycoplasma mycoides clade**. *Int J Syst Evol Microbiol* 2004, **54**(3):893-918.

144. Mushegian A: **The minimal genome concept**. *Curr Opin Genet Dev* 1999, **9**(6):709-714.

145. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes**. *Proc Natl Acad Sci U S A* 1996, **93**(19):10268-10273.

146. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, 3rd, Smith HO, Venter JC: **Essential genes of a minimal bacterium**. *Proc Natl Acad Sci U S A* 2006, **103**(2):425-430.

147. French CT, Lao P, Loraine AE, Matthews BT, Yu H, Dybvig K: **Large-scale transposon mutagenesis of Mycoplasma pulmonis**. *Molecular microbiology* 2008, **69**(1):67-76.

148. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Debarbouille M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan R, Eschevins C *et al*: **Essential Bacillus subtilis genes**. *Proc Natl Acad Sci U S A* 2003, **100**(8):4678-4683.

149. Yashina S, Gubin S, Maksimovich S, Yashina A, Gakhova E, Gilichinsky D: **Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost**. *Proc Natl Acad Sci U S A* 2012, **109**(10):4008-4013.

150. Huynen MA, Bork P: **Measuring genome evolution**. *Proc Natl Acad Sci U S A* 1998, **95**(11):5849-5856.

151. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W: **Reconstructing contiguous regions of an ancestral genome**. *Genome Res* 2006, **16**(12):1557-1565.

152. Tuller T, Birin H, Gophna U, Kupiec M, Ruppin E: **Reconstructing ancestral gene content by coevolution**. *Genome Res*, **20**(1):122-132.

153. Manso-Silv √ ° n L, Vilei EM, Sachse K, Djordjevic SP, Thiaucourt F, Frey J: **Mycoplasma leachii sp. nov. as a new species designation for Mycoplasma sp. bovine group 7 of Leach, and reclassification of Mycoplasma mycoides subsp. mycoides LC as a serovar of Mycoplasma mycoides subsp. capri**. *International Journal of Systematic and Evolutionary Microbiology* 2009, **59**(6):1353-1358.

154. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families**. *Science* 1997, **278**(5338):631-637.

155. Darling AC, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements**. *Genome Res* 2004, **14**(7):1394-1403.

156. Tesler G: **GRIMM: genome rearrangements web server**. *Bioinformatics* 2002, **18**(3):492-493.

157. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

158. Grishin NV, Wolf YI, Koonin EV: **From complete genomes to measures of substitution rate variability within and between proteins**. *Genome Res* 2000, **10**(7):991-1000.

159. Felsenstein J: **PHYLIP -- Phylogeny Inference Package (Version 3.2)**. *Cladistics* 1989(5):164-166.

160. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W182-185.

161. Darling AE, Mau B, Perna NT: **progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement**. *PLoS One*, **5**(6):e11147.

162. Darling AE, Miklos I, Ragan MA: **Dynamics of genome rearrangement in bacterial populations**. *PLoS Genet* 2008, **4**(7):e1000128.

163. Larget B, Simon DL, Kadane JB, Sweet D: **A bayesian analysis of metazoan mitochondrial genome arrangements**. *Mol Biol Evol* 2005, **22**(3):486-495.

164. Bourque G, Pevzner PA: **Genome-scale evolution: reconstructing gene orders in the ancestral species**. *Genome Res* 2002, **12**(1):26-36.

165. Oshima K, Nishida H: **Phylogenetic Relationships Among Mycoplasmas Based on the Whole Genomic Information**. *Journal of Molecular Evolution* 2007, **65**(3):249-258.

166. Brown DR, Whitcomb RF, Bradbury JM: **Revised minimal standards for description of new species of the class Mollicutes (division Tenericutes)**. *Int J Syst Evol Microbiol* 2007, **57**(Pt 11):2703-2719.

167. Davidson AL, Dassa E, Orelle C, Chen J: **Structure, function, and evolution of bacterial ATP-binding cassette systems**. *Microbiol Mol Biol Rev* 2008, **72**(2):317-364, table of contents.

168. Chambaud I, WrÛblewski H, Blanchard A: **Interactions between mycoplasma lipoproteins and the host immune system**. *Trends in Microbiology* 1999, **7**(12):493-499.

169. Bennett PM: **Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement**. *Methods Mol Biol* 2004, **266**:71-113.

170. Pilo P, Frey J, Vilei EM: **Molecular mechanisms of pathogenicity of Mycoplasma mycoides subsp. mycoides SC**. *Vet J* 2007, **174**(3):513-521.

171. Halbedel S, Hames C, Stulke J: **Regulation of carbon metabolism in the mollicutes and its relation to virulence**. *J Mol Microbiol Biotechnol* 2007, **12**(1-2):147-154.

172. Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, Arcus VL: **On the origin and evolution of thermophily: reconstruction of functional precambrian enzymes from ancestors of bacillus**. *Mol Biol Evol*, **29**(2):825-835.

173. Zwickl DJ: **Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion**. *PhD dissertation, The University of Texas at Austin*
2006.

174. Sukumaran J, Holder MT: **DendroPy: a Python library for phylogenetic computing**. *Bioinformatics*, **26**(12):1569-1571.

175. Yang Z, Nielsen R, Hasegawa M: **Models of amino acid substitution and applications to mitochondrial protein evolution**. *Mol Biol Evol* 1998, **15**(12):1600-1611.

176. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods**. *Molecular Biology and Evolution*, **28**(10):2731-2739.

177. **The PyMOL Molecular Graphics System** In., Version 1.3 edn: Schrödinger, LLC.

178. Suzuki Y, Glazko GV, Nei M: **Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics**. *Proceedings of the National Academy of Sciences* 2002, **99**(25):16138-16143.

179. Alcaraz LD, Moreno-Hagelsieb G, Eguiarte LE, Souza V, Herrera-Estrella L, Olmedo G: **Understanding the evolutionary relationships and major traits of Bacillus through comparative genomics**. *BMC Genomics*, **11**:332.

180. Tsuchiya D, Sekiguchi T, Takenaka A: **Crystal structure of 3-isopropylmalate dehydrogenase from the moderate facultative thermophile, Bacillus coagulans: two strategies for thermostabilization of protein structures**. *J Biochem* 1997, **122**(6):1092-1104.